SIXTH FRAMEWORK PROGRAMME

PRIORITY 2

Information Society Technologies

Project n°033841

EMIL

Emergence In the Loop: Simulating the two way dynamics of norm innovation

Deliverable 2.2

Report on norm innovation main features

Due date of the Deliverable: 31 Ago 08 + 45 dd.

Actual Submission date: 1 Oct 08

Start date of the Project: 01 Sept 2006 Duration: 36 months

Organization responsible for this Deliverable: UNIS

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Index

Explaining Normative Behavior in Wikipedia

Introduction

The University of Surrey, Centre for Research on Social Simulation is responsible for identifying cases and for the conduct of primary and secondary research into those cases to provide empirical data to inform both theory development and simulator design. The studies conducted as part of this work program are to be informed by and to inform the ontology (EMIL-M) and to provide data as a point of comparison for the simulator (EMIL-S). The intention was also to bring a social science perspective to the study of normative behaviour in the selected cases. Work undertaken in 2008 included:

- Extension of the preliminary empirical research into Wikipedia reported in 2007 and considering normative influence on Discussion pages in both Controversial and Featured articles.
- Further development of a theory of orders of emergence.
- Identification of hypotheses which might guide further simulation and empirical research into normative behaviour.

This work has been developed through presentation to and participation in a number of workshops and conferences and in published form as follows.

- Goldspink, C. & Kay R. 2009, 'Autopoiesis and organizations: A biological view of organizational change and methods for its study'. Forthcoming in, Magalhaes, Rodrigo, Sanchez, Ron (eds), "Autopoiesis in Organizations and Information Systems", Elsevier Science (Advanced Series in Management).
- Goldspink C. 2009, 'Social Self Regulation in On-line Communities: The Case of Wikipedia', International Journal of Agent technologies and Systems 1(1).
- Goldspink C. & Kay R, 2009, 'Agent Cognitive Capability and Orders of Emergence', forthcoming in Trajkovski, Goran and Collins, Samuel (eds), "Agent-Based Societies: Social and Cultural Interactions"
- Goldspink, C. Edmonds, B. & Gilbert N. 2008, 'Normative Behaviour in Wikipedia', paper delivered at the e-social Science Conference, Manchester, June 18-20
- Goldspink, C. & Kay R., 2008, 'Agent Cognitive Capabilities and Orders of Emergence: critical thresholds relevant to the simulation of social behaviours', In proceedings of AISB Convention, Communication, Interaction and Social Intelligence Aberdeen, UK April 1-4

Other contributions of particular relevance:

- Symposium Chair, Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB) 2008 Conference, Aberdeen, Scotland UK.
- Participant in 'Enactive Approaches to Social Cognition Workshop' Battle, UK, Aug 30th – Sep 1st, 2008, a Sixth Framework Programme - Information Society and Technologies - Citizens and Governance in the Knowledge Based Society funded initiative.

Wikipedia Case Analysis

When people encounter Wikipedia for the first time and learn how it works, they commonly express surprise. The expectation appears to be that an open collaborative process of such magnitude could not work. Yet Wikipedia has been shown to produce credible encyclopaedic articles (Giles, 2005) without the hierarchical and credentialist controls typically employed for this type of production. Our interest is with the origin and form of order which underpins these results and the role and mechanisms of normative influence which they may reveal.

In the course of 2008 we considered this question from the perspective of Governance theory (Goldspink 2009). The governance model which arguably comes closest to explaining the phenomenon of Open Source production in general has been called Bazaar Governance. With Bazaar Governance (Demil & Lecocq, 2003; Raymond, 2001) there is no obligation on any party to perform particular duties or even to remain engaged: there are low entry and exit costs. There are few formal mechanisms for policing or for sanction but sufficient regulation is achieved by means of shared task, reciprocity norms and/or informal group sanctioning with participants influenced by their desire to build reputation. However the limited effect of reputation in Wikipedia suggests that it may be better considered through the more general lens of stewardship theory. Norms are argued to play a role in both Stewardship, and Bazaar theories of governance and this aligns this research with the needs of the EMIL project.

However, while sociologists have long argued that norms are fundamental mechanisms for social regulation there is no entirely convincing answer to the question "what is a norm?"

In the social science literature norms are approached in two contrasting ways. These may be summarised as *constitutive* or *regulative*. The former is associated with the social philosophical tradition (Lewis, 1969). Here norms are seen as a particular class of emergent social behaviour which spontaneously arise in a population: a 'norm' is a pattern identified by an observer ex-post. The defining characteristic is the apparently prescriptive/proscriptive character: people behave 'as if' they were following a rule. By contrast, the regulative view derives from the philosophy of law. Here a norm is seen as a *source* of social order. This standpoint assumes the prior existence of (powerful) social institutions and posits them as the source of rules, which, when followed, lead to social patterns. These positions appear antithetical although following the work of Berger and Luckman (1972) each may be seen as a part of a dialectic whereby emergent social patterns become internalized at the level of individuals and formalised in institutions.

In this research we have avoided making a prior commitment to either of these approaches and instead have pursued an understanding which is both empirically grounded and which attempts to explicate the dialectical relationship between social structure (as exemplified by the regulative view) and individual agency and collective interaction (as exemplified by the constitutive view). Our purpose is to explore the possible origins and mechanisms of social order in this particular on-line social system. We do this by examining how the various attributes of the system might combine to explain differences in the quality of Wikipedia articles and the apparent role of 'norms' in the attainment and maintenance of quality.

Analysis of Self-Regulation mechanisms

In Wikipedia there are two classes of activity: editing; and conversation about editing. Our work so far has not been directly concerned with the editing activity (although this is to be considered in future research) but with the conversations which help to coordinate it. The quality of the final article is influenced by the collective consequences of editing activity: this is dynamic as a page is never final. The editing activity *may* be influenced by the conversations had between editors and by administrative actions that result from these discussions (such as locking an article from further editing for a time or blocking a user). Insight into the way conversation does or does not influence the final product can be gained by examining the Discussion and Talk pages. In this study we focused on the Discussion

pages. These contain the most 'public' of the discussions and these pages are expressly established to provide a forum for discussions about specific articles: their content and presentation.

The activity on Discussion pages comprises a series of 'utterances' or speech acts between contributors about editing activity and the quality of product. On the face of it, these pages should provide a fertile source of data to support analysis of how social regulation works in the Wikipedia, in particular 'soft' regulation. Within these pages we expected to see attempts by editors to influence the behaviour of one another through the only means available to them – communicative acts. We anticipated that these may exhibit some regularity which would allow us to examine both the range and type of events that led to the explicit invocation of rules and norms and which revealed emergent influence patterns which were themselves normative. We wanted also to examine what conventions prevailed and how these compared and interacted with the goal of the community and its policies. A convention is defined here as a behavioural regularity widely observed by members of the community. Policies include explicit codes of conduct as well as guidelines (etiquettes) and principles.

Method

For the study we randomly selected a sample of Discussion pages associated with both Controversial and Featured articles. At the time of the study (May/June 2007) there were 583 articles identified by the Wikipedia community as controversial and approximately 1900 as featured. The analysis reported here is based on a sample of 19 Controversial and 11 Featured articles. The most recent three pages of discussion were selected for analysis from each Discussion associated with the article included in the sample. These were subjected to detailed coding using the Open Source qualitative analysis software WeftQDA. Both qualitative and quantitative analysis was performed. The latter was undertaken by re-processing the coded utterances such that each utterance constituted a case and each applied code became a variable associated with that case. This data was then analysed using SPSS and MLwin.

A number of coding schemes for natural speech were considered before choosing the Verbal Response Mode (VRM) taxonomy (Stiles, 1992). VRM has been developed over many years and used in a wide range of communication contexts. Stiles defines it as 'a conceptually based, general purpose system for coding speech acts. The taxonomic categories are mutually exclusive and they are exhaustive in the sense that every conceivable utterance can be classified.' (Stiles, 1992: 15). The classification schema is attractive where there is a need (as here) to capture many of the subtleties of natural language use that derive from and rely on its intrinsic flexibility and ambiguity yet map them to a more formal system needed for analysis. Additional codes were applied to identify: valence, subject of communication, explicit invocation of norms or rules and the associated deontic and trigger, whether the receiver/s accepted the illocutionary force of the utterance and the ID and registration status of the person making the utterance. There were 3654 utterances coded in these thirty three documents.

Findings

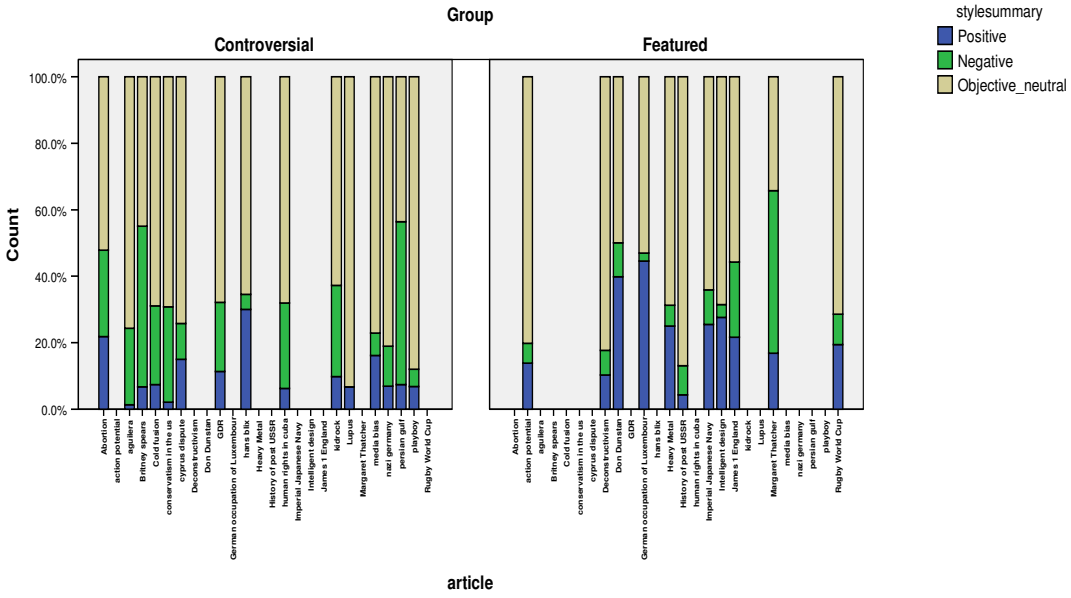
Style of Communication

There was a statistically significant correlation between the article group (Controversial *vs* Featured) and broad style of communication. This was however very small at -0.078 ($p < 0.01$, 2-tailed). This difference was most apparent when examined at the level of specific styles. Both groups had approximately similar proportions of neutrally phrased utterances (approximately 64%). Nearly one quarter (22.5%) of all utterances in the discussion pages of Featured articles were positive compared to only 11% in those of Controversial ones. By comparison nearly one quarter (23.9%) of all utterances in Controversial discussion pages

were negative compared to 14% for featured. The positive styles of 'affirming', 'encouraging' and 'acknowledging' were significantly overrepresented in the discussion pages of Featured articles but underrepresented in the Controversial ones. The reverse was the case for the negative styles of 'aggressive', 'contemptuous' and 'dismissive' in the controversial. Overall, the most common positive utterance was affirming (4.7%) closely followed by encouraging (4.7%) and acknowledging (4.3%). The most common negative utterance was dismissive (8.2%) followed by defensive (6.4%) and contemptuous (3.5%). All the Wikipedia discussions sampled reflected a strongly neutral-objective *style* (although it was apparent from the Qualitative study that the content was sometimes far from objective or balanced).

The broad style variable was coded such that it formed an ordinal variable comprising three states, low, medium and high which were dummy coded to 1, 2 and 3 respectively. The distribution was approximately normal with a mean of 1.96, an SD of 0.6 and n=3575. This dummy coded variable was used to further examine the sources of variance. A one way ANOVA of this variable by group (Controversial vs Featured) confirmed that there was a statistically significant difference (p<.000) but with small effect size (Eta squared .03). Between group variance Sum of Squares was just 38.8 compared to a within group Sum of Squares of 1228.2 (F = 112.8). The high level of variation between articles within each group is apparent in the following chart.

Chart one: Percentage of style by article by group.

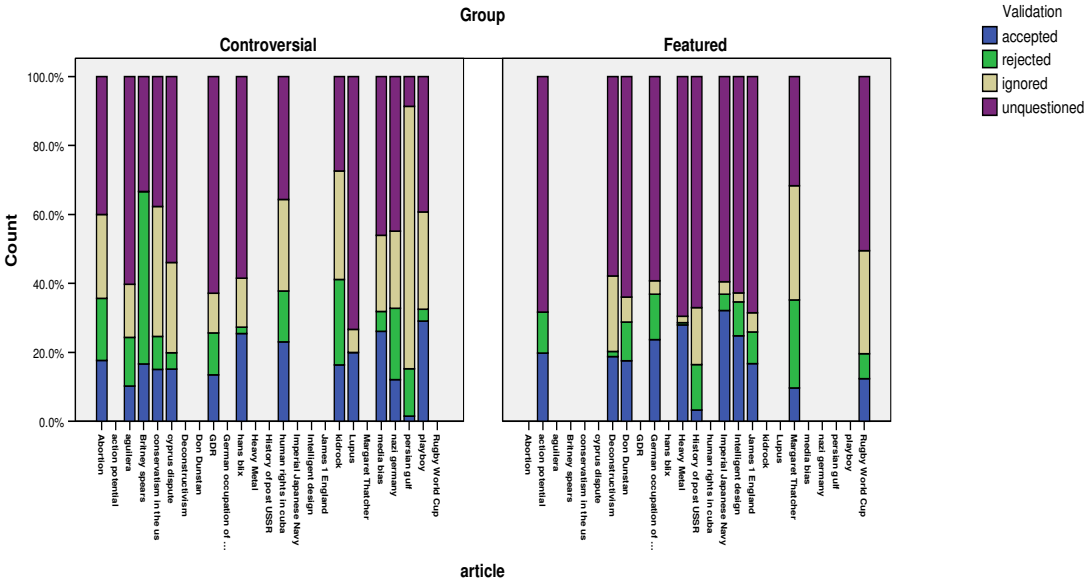


In order to better understand the origin of this variance between articles a simple multilevel analysis was performed. The sample used was not specifically designed for MLA, however analysis showed that it did satisfy the normally accepted sample minimums (Maas & Hox, 2005). Three levels were identified as possible sources of variance: article (n=30), thread (n=233) and utterance (n=3654). The communication style variable was reprocessed to provide a simple three interval scale with the values of low (-1) neutral (0) high (+1). While a three interval scale is far from ideal it was anticipated that it might prove sufficient for providing at least some indication of where the primary variance was to be found. This analysis showed coefficients at the level of the article estimated at .011 (p<.01), at the level of the thread of .07 (p=.01) and at the level of the utterance of .27 (p<.01). This suggests that differences in communication style have very little to do with differences in the articles itself and only a little to do with the specific topics of discussion.

Validation

Within speech act theory (Habermas, 1976; Searle, 1969), validation refers to whether an utterance made by one speaker is accepted, rejected, ignored or let go unquestioned by the intended recipient/s. In the Wikipedia sample 50% of all utterances were accepted without question. A further 18% were explicitly accepted by at least one editor; 11% were explicitly rejected and a substantial 22% were ignored. 25% of positive style utterances were accepted by at least one editor compared to 18% of neutral and only 9% of negative. By comparison only 2% of positive utterances were rejected compared to 9% of neutral and 26% of negative. Positive utterances were more likely to be accepted without question (61%) compared to negative (21.7%) and neutral (54.4%). Negative comments were more likely to be ignored (44.1%) compared to neutral (18.2%) and positive (11.4%). From this we can conclude that positive utterances are more likely to be validated than negative, but that overall, a significant number are ignored or rejected. A similar procedure of forming an ordinal variable from the four scale items was undertaken. Here reject was recoded to 1 as the strongest form of rejection, followed by ignore, unquestioned and accept. Again the resulting distribution was approximately normal, with a mean of 2.7 SD of .88 N=3148. The ANOVA confirmed a statistically significant between group difference (p<.000). The Between group sum of squares was 47.3 and the Within group sum of squares 2373.5 resulting in a small Eta squared of .02 (F=62.7).

Chart two: Percentage of validation by article by group.



Normative and rule invocation

A norm is coded where an editor specifically invokes a norm which is a) not the subject of an existing Wikipedia rule or b) the editor does not link to a specific Wikipedia practice or rule (even if one exists) but rather refers to a wider social standard. A rule is coded where an editor explicitly cites a rule present in a Wikipedia guideline, etiquette or style guide. Overall 5.2% of all utterances involved norm or rule invocation. This meant that Wikipedia rules were invoked 122 times and general social norms a further 77 times in 3654 utterances. This overall number was contributed to disproportionately by three (outlier) articles in the sample. Without these outliers the rate of invocation was in the order of 2%. Rules were most commonly invoked in response to neutral style communication (63.9%) followed by 27% in

response to a negative style. Only 9% of positive style utterances were responded to with a rule invocation. By comparison, norms were most commonly invoked in response to negative style utterances (53.2%) followed by neutral (44.2%) and then positive (2.6%). The difference in likelihood of invocation by style was statistically significant (p=.001).

A Wikipedia rule invocation was most likely to be triggered by the *form* of an article (44.9%), an *edit action* (22%), an *article fact* or a *person's behaviour* (both 16%). A norm was most likely to be triggered by a *person's behaviour* (35.6%), an *edit action* (23.3%), *article form* (21.9%), or *article fact* (19.2%). This pattern did not differ to a significant degree between the Featured and Controversial articles. Nearly three quarters (73.6%) of rule invocations had the implicit deontic of 'it is obligatory'. Norms also were most likely to carry this deontic (61.3%). The second most likely deontic was 'it is permissible that' (9.7%).

While there was no statistically significant difference in the degree to which either norms or rules were invoked between the Featured and Controversial articles, there was a qualitative difference in the role norm and rule invocation appeared to play. In Controversial discussions, social norms and rules were most likely to be invoked against the behaviour of an editor who was of a different view while in Featured sites, norms and rules were somewhat more often used by the editor as a reflection on their own contribution – i.e. involved a level of self-check.

Registered vs non-registered users

Although there was no statistically significant difference in the likelihood for either registered or non-registered editors to invoke norms or rules, there was a statistically significant difference between registered and non-registered editors ($p < 0.01$) when it came to validation. Registered editors were more likely than non-registered to be explicitly accepted (18.7% of utterances compared to 13.9%), less likely to be rejected (9.9% compared to 13.7%), considerably less likely to be ignored (18.3% compared to 34.7%) or unquestioned (53.1% compared to 37.6%). Qualitatively, however, it was much more common that un-registered users would make suggestions before undertaking edits, particularly in the Featured articles, so their behaviour was less likely to attract action or comment. Non-registered editors were more likely to make negative style utterances (24.3% compared to 18.5%) and less likely to make positive utterances (9.5% compared to 17.4%). This difference was significant ($p < 0.01$).

Influence through Illocutionary Force

The theory of speech acts distinguishes between the meaning of an utterance and its pragmatic intent. A typical utterance may have a *form* that differs from the *intent*. The utterance 'could you close the door?', for example, has the form of a *question* but the intent of *advisement:* the speaker intends the listener to close the door. With the VRM coding frame used in this research each utterance is coded twice, once to capture the semantic form and again to capture the use of language to exert (illocutionary) force (Searle, 1969). In VRM, the relationship of form to intent is expressed using the statement "in service of" (Stiles, 1992). In this example the question 'could you close the door' is 'in service of' the advisement 'close the door'. In standard presentation this is recorded as (QA).

Edification in service of Edification (EE) is the most frequent form of utterance in the Wikipedia sample – 37% of all utterances were of this mode. The Edification mode is defined as deriving from the speaker's frame of reference, making no presumption about the listener and using a neutral (objective) frame of reference shared by both speaker and listener. This mode is informative, unassuming and acquiescent. It reflects attempts to convince by neutral argument. An example would be 'That edit was made last week'.

The second most common mode is that of Disclosure in service of Disclosure (DD). Disclosure is defined as being from the speaker's experience, making no presumption, but being framed using the speaker's frame of reference. This is summarised as informative,

unassuming but directive. Unlike EE mode, DD mode represents an attempt by the speaker to impose or have the listener accept the speaker's frame. 12% of all utterances adopted this form. An example would be 'I don't know much about this topic'.

The third most common mode is Disclosure in service of Edification (DE). The DE mode represents an utterance which is from the speaker's frame of reference but as if it is neutral or from a shared frame. 8% of all utterances used this mode. This is a somewhat neutral mode where the speaker offers clearly labelled personal knowledge as information. An example would be 'I believe it occurred in 1987'.

The fourth most common mode is Advisement in service of Advisement (AA). AA mode represents speech from the speaker's experience, which makes presumptions about the listener and adopts the speaker's frame of reference. It can be summarised as informative, presumptuous and directive. An example would be 'You should change this immediately'. Approximately 7% of utterances were in this mode. A further 12% of utterances have the directive pragmatic intent of advisement masked by a less presumptuous form – Edification or Disclosure ('It should be changed immediately' or 'I think it should be changed immediately').

Significantly, utterances associated with politeness were very rare in this sample.

Discussion of Findings

What is significant about the utterance strategies is that they typically involve an exchange of assertions delivered with a neutral – i.e. non-emotive style. There are very few explicit praises or put downs, and few niceties like explicit acknowledgements. Seldom do contributors refer to one another by ID – the exchanges are very impersonal. This does not tally with what one would expect if the Wikipedia etiquette (http://en.wikipedia.org/wiki/Wikipedia:Etiquette) had been institutionalised. The Featured articles conform a little more closely with what one would expect than do the Controversial, but if we assume that the etiquette captures the community's ideal, the utterances do not conform to that 'ideal' in either case.

For example the Etiquette specifically encourages politeness and reminds editors that the lack of emotional cues is an issue in e-communication. An excerpt states:

Give praise when due. Everybody likes to feel appreciated, especially in an environment that often requires compromise. Drop a friendly note on users' talk pages.

- *Be open and warmly welcoming, not insular, Make others feel welcome (even longtime participants; even those you dislike),*
- *Create and continue a friendly environment,*
- *Turn the other cheek (which includes walking away from potential edit wars),*
- *Give praise, especially to those you don't know (most people like to know they are wanted and appreciated), and*
- *Forgive!*

Similarly we see low levels of questioning or of reflection (i.e. feeding back the words of the speaker to check understanding or to come to a better understanding of the other's intentions). This is inconsistent with the task needs – to reach consensus on controversial topics. The frequency with which utterances were ignored also suggested low engagement by participants in the discussion. All of this would seem to need some explanation.

The absence of any expression of acknowledgement of emotions and/or similarity of attitude (homophily) among many contributors suggests that Wikipedia lacks many of the qualities of verbal exchange that would identify it as strong community. These signals do appear to be more common on personal Talk pages but here it appears to be in the context of in-group 'stroking', that is associated with individuals who find they are in agreement reinforcing their

sense of belonging to a common group. The interaction on Discussion pages, by contrast, is more consistent with being a place to share coordination of a task – like a work place formal meeting where the participants are relatively unknown to one another and where trust is not assumed. This could suggest that the goal is the primary orientating point. However, the lack of quality of discourse needed to achieve consensus is more indicative of a brief encounter between individuals who struggle (or are not fully committed) to find common understanding rather than of a community committed to a common goal (Becker & Mark, 1997). This might suggest that the shared goal may be subordinate to more personal goals for a considerable proportion of contributors. Or it may be that the technology and environment will support no more than this.

The Wikipedia environment supports saboteurs who can use the opportunity afforded by the open and anonymous platform to use identity deception i.e. to mimic the language and style of an 'expert' or to present as a genuine editor while trying to pursue a personal or political agenda hostile to the aims or interests of the Wikipedia. We found no direct evidence of this behaviour in the Discussion pages we sampled even though the discussions about controversial articles provide particularly fertile ground for such sabotage. However, vandalism is not generally evident in Discussion pages only on articles, there were some interactions which could have been regarded as trolling but were not explicitly labelled as such. Nevertheless the threat of these behaviours could have an influence on the type of communication conventions which arise. Editors may, for example, display reserve and suspicion, withholding trust and taking conventional signals of authority and identity (Donath, 1998) as unreliable.

Qualitatively there was considerable evidence that editors appeared to form judgements about the intent of others on relatively little information. There was, however, little evidence of the use of utterance strategies to better understand or check these judgements. Only occasionally would an editor modify his/her style significantly if challenged. Of the rule invocations 26% were accepted, a similar proportion were rejected or ignored and the remainder went unquestioned. This is consistent with norms being triggered by a limited range of cues which allow individuals to locate themselves and select identities appropriate to a context and which then remain essentially stable. The invocation of rules and norms appears to have little to no immediate effect on behaviour. It is not clear if it has an effect on future behaviour as this cannot be ascertained from the available data.

Conclusions

In this study we set out to identify the mechanisms that underpin the emergence of systemic self-organisation in a volunteer on-line global institution. The findings have challenged some of our assumptions and expectations, in particular:

- The detailed and specific behavioural etiquette published in Wikipedia seems to have little influence on the overall character and style of interaction on Discussion pages.
- The overall quality of interaction of editors on Discussion pages falls short of the range and quality of communicative style characteristic of a community and also of that which would be expected, given the nature of the task.
- Most regulation is achieved without the need for frequent explicit invocation of rules or norms. Rather, behaviour seems to accord to a convention which editors quickly recognise and conform to (or bring to the Wikipedia) and which minimally accommodates what needs to be done to satisfy the task in a context of divergent personal goals.
- There was a lack of evidence of active negotiation of expectations and standards and convergence of behaviour towards a norm. Within the discussion pages there appeared to be little obvious norm innovation, evolution, adaptation or extension. This suggests that on first encounter with Wikipedia, editors read a set of cues as to

what constitutes appropriate or acceptable behaviour and then more or less accommodate to it. Alternatively the order observed may be largely attributable to the prior socialisation of participants with local norms and rules playing a very minor role.

- While there is a difference between Controversial and Featured article Discussion pages this is small and the quality of the interaction cannot explain the difference in status. Similarly there appeared to be little in the subject matter of the two groups of articles that would explain the difference – both contained subject matter which was contestable and subject to significantly diverse opinion.
- There is no clear basis to argue that the apparent order is a direct result of the use of deontic commands associated with social norms and environment specific rules. Despite the fact that the community has been a prolific rule generator, they appear to play a minor role. Contributors demonstrate a style which is broadly inconsistent with these rules and not a good fit with the task.
- Overall though there is order and it appears to be emergent. The mechanisms that underpin this emergence have not been revealed by the analysis undertaken to date although some hypotheses can be tentatively suggested. The neutral-objective style may be a consequence of the anonymity and open nature of the environment – leading to a suspension of trust. It may propagate as newcomers copy the pattern through a process of behavioural cueing. It is possible also that the order is due to pro-social behaviour internalized and brought to the task. The volunteer nature of Wikipedia, and the level of commitment required, is likely to mean that long term editors reflect a pro-social disposition (Penner et al., 2005b). In this context a little norm/rule invocation may go a long way, if not by influencing immediate behaviour, then by encouraging future compliance and/or by giving an incentive for non-compliers to leave. Such a view is quite different from that presumed by previous theories of social norms.

While the findings of the research to date are far from conclusive they do challenge many of previous assumptions and suggest a range of alternative hypotheses. Some of these will be able to be examined by further analysis of the current data and/or by data currently being collected through a controlled wiki experiment as well as data proposed to be collected into the relationship between discussions and editing behaviour. These should allow us to test alternative hypotheses and contribute to our understanding of how social regulation works in highly open computer mediated environment.

Formulating Testable Hypotheses

In order to support systematic exploration of the issues identified through the empirical work hypotheses need to be presented which frame the issues in a manner which supports simulation based experiments. Not all of the issues identified lend themselves to experimentation using social simulation but suggest the need for further conventional social research before the problem can be specified with sufficient clarity to implement in the form of a simulation. To help frame the different hypotheses using a similar framework activity theory is used.

The Wikipedia as computer mediated social system

Wikipedia is a computer mediated workspace with very distinctive technical and social characteristics. One approach that has been enjoying considerable application for the study of computer mediation is activity theory. In Activity theory, unsurprisingly, **activity** is the unit of analysis. An **activity** is a complex process, below it sit: **actions** – specific tasks executed in a time bound manner in order to achieve the object; and routine or automatic **operations**, a particular act in a time and place which does not have an independent goal but serves to adjust

an action to a current situation (Kaptelinin & Nardi, 1997). In the case of Wikipedia the activity is the development of a quality article. This involves editors in actions such as reducing the 'Point Of View', or improving the Grammar of an article as well as having conversations about these issues on Discussion and Talk pages. Finally we have operations – specific edits and utterances about edits.

The trajectory of an activity system is described as 'far from equilibrium' driven by tensions and contradictions between aspects of the activity system itself. The following figure shows the essential elements of such a system.

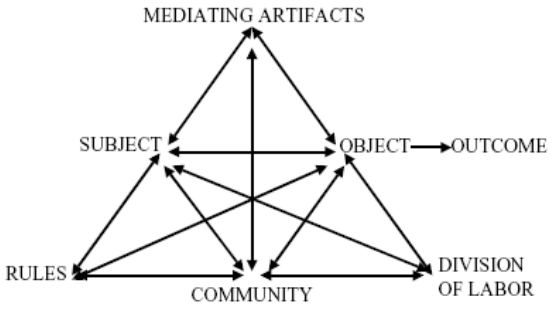


Figure 1: Engeström's Activity System Model [Engeström 1999]

Activity theory recognizes two basic processes - internalization and externalization. These are argued to operate continuously at every level. According to Engestom (1999: 10) internalization is related to the reproduction of culture. In the language used within the multi-agent modeling community this is the process of *immergence* (Castelfranchi, 1998a). Immergence can be thought about in a number of different ways: as Castelfranchi conceived of it, drawing on the perspective of First generation AI or cognitivism, it involves a 'representation' in the mind of the agent. From an enactive perspective it involves change in the nervous systems which alters future dispositions for action. These changes at the level of participating agents influences the way they act so as to maintain or alter cultural patterns through the process of emergence. Externalization is the trace of cultural pattern embedded in tangible or symbolic artifacts – emergent form made concrete or rendered symbolic. Activity shapes artifacts and the particular characteristics and form of artifacts subsequently shapes the interaction of agents and becomes internalized as mental activity. These artifacts or tools *'…connect an individual to other human beings by mediating activity, thereby becoming part of a cultural context'* (Fjeld et al., 2002: 157). It is the focus on the potentially transformative effect of mediating artifacts that has led to a resurgence of interest in activity theory, most notably in the computer sciences. Here it is seen as a relevant framework for examining human computer interaction (Fjeld et al., 2002; Kaptelinin & Nardi, 1997; Lewis, 1997) as well as to gain insight into the design of multi-agent systems (Ricci et al., n.d). It is cast not as a predictive theory but rather as a meta-theory and that is how it is used here. Activity theory is useful for identifying specific hypothetical relationships between aspects of an activity system which may explain observed behavior or outcomes. More conventional methods will generally be needed to collect evidence to support any specific hypothesis.

Applying activity theory to Wikipedia

Mediating Artifacts

The artifacts relevant to understanding Wikipedia include:

- The distinctive features of the WikiWiki technology platform and its associated affordances and sub-systems such as bots as well as of other technologies used by editors such as email and list servers;
- The genre of Encyclopedic writing;
- Written form of presentation and organization of rules and guidelines;
- Patterns in the text of existing article Discussion and Talk pages.

Considering each of these in turn.

WikiWiki technology has a **very flat learning curve**: contributing is extremely simple. There are few technical impediments confronting novice users. Wiki platforms are **intrinsically open** supporting decentralised action unless modified to control or restrict access. Wikipedia has added a number of facilities which support the ready detection and correction of vandalism. **Watch lists** support users in taking responsibility for the oversight and monitoring of particular topics. Changes made to a page are logged using a **history list** which supports comparison between versions as well as identifying the time and date of any change and the ID of who made that change. The **reversion** facility supports the rapid reinstatement of the page content. D**iscussion pages** which support coordination, negotiation and alignment of individuals views about article form and content. Ciffiolilli (2007) has argued that a significant consequence of these technical features is that they reduce transaction costs associated with complex coordination. However, the technology does not cancel other costs of coordination and control. These are commonly referred to as agency costs. In Wikipedia they are borne by individual editors and not necessarily equitably. The agency cost burden will be less where there is a high level of self-regulation and lower where a lack of goal alignment or low social commitment leads contributors to disregard others and act individualistically or opportunistically.

The Encyclopedia Genre: everyone knows what an encyclopedia is and how it is organized. This is a powerful artifact which serves as a backdrop to all editors contributions. Its importance as an artifact has been noted by the founding editor Larry Sanger (2005) and is reflected in the 'What Wikipedia is Not' pages and discussions.

Patterns of Discussion Pages: Discussion pages will generally have a history and the focus, style and tone of past conversations may influence how communication proceeds when others add to it.

Subjects

The subjects are the individual editors associated with a particular article. This is an area where we lack information. As Wikipedia editors remain anonymous, it is difficult to collect data about them or from them as individuals and the few available studies generally involve very small samples (see for example Forte & Bruckman, n.d) What is particularly missing from existing research is information on basic demographics as well as about editors motivation to invest the time to contribute in the first place and to stay involved despite the high agency costs they may face.

Rules,

There are many rules and norms relevant to all aspects of article development, framing, style and content as well as the way in which editors behave. Rules may include those specific to the Wikipedia or relevant in wider social groups that editors operate within. They may be written or they may be 'generally understood'.

The community

This refers to the wider social grouping or 'community of practice', to which the subject belongs: the Wikipedia community. This is somewhat more problematic as in principle anyone with access to the Internet can be in this community, there are no rules or requirements for membership nor any obligations of commitments to remain. That said, there do appear to be a core of relatively stable contributors associated with particular articles. In their study of who contributes to Wikipedia's value, Priedhorsky et al (2007) found that the editors who edit many times dominate what people see. The top 10% of editors measured by number of edits contributed 86% of their persistent word view score which measures the number of times any word they contributed is viewed by others. The top 0.1% of editors contributed 44% of persistent word views. Furthermore this proportion of value created by a few is increasing over time. Clearly this is important work for some and they must dedicate considerable time to the endeavor where others are content to make a passing spelling correction. Elsewhere we have raised the question about whether the community is deserving of the title community (Goldspink & Kay, 2009). The absence of any expression of acknowledgement of emotions and/or similarity of attitude (homophilly) among many contributors suggests that Wikipedia lacks many of the qualities that would identify it as strong community. Possibly it therefore fails to constitute a distinct domain of discourse. It may be better considered as a place to share coordination of a task with a shared goal as the primary orientating point. However, even here, the lack of quality of discourse needed to achieve consensus is more indicative of a brief encounter between different and established milieux which struggle to find common understanding rather than of a community committed to a common goal (Becker & Mark, 1997). Forte and Bruckman (2008) have recently noted that relationships and a capacity to influence appears to be associated with three categories of user: unregistered, registered and Arbitration Committee member and that the structure is increasingly becoming decentralized with local differences emerging. There are therefore clearly some boundary issues to be come to terms with in considering the Wikipedia community. One perspective is to see it as a coalition of loosely coupled groups (Orton & Weick, 1990).

Perhaps the area where 'the community' has greatest impact is on voting for sites to be classified and electing people to administrative roles. It shapes the flagging of problem or exemplar artifacts and impacts also on the division of labor.

Division of labor

The division of labor occurs in several ways and using a variety of mechanisms. The first form of division is with formed when an editor chooses to be associated with particular articles: this is self-initiated and changeable. The work of (Brandes et al., 2008) has shown that within the group of editors associated with an article there emerges a distinction between those who choose to make substantive contributions and those who defend the article from vandalism. In some cases there are further divisions around topic with defenders or advocates of particular positions and mediators emerging (polarization), the balance of these may have important implications for the resulting quality, particularly in terms of neutrality. Finally there is the division which arises with the creation and election of individuals to positions such as Administrator and Bureaucrat or to roles such as mediator and arbitrator.

Object

The **object** is the article or articles which together make up Wikipedia. These are themselves key artifacts – embedding the characteristics of the genre. The form of the articles is a primary source of orientating feedback about how well the activity has or has not succeeded in achieving the goal.

Hypotheses

From a social science perspective, and in particular from the perspective of the EMIL project, with its focus on understanding normative regulation, a number of aspects of empirical interest emerge from prior research into Wikipedia. These include:

- The differences in **quality** between Featured and Controversial Sites: is this a product of some form of normative self-regulation that perhaps works differently in relation to these different articles resulting in different quality attributes in the object?
- In the context of Wikipedia what is the role of explicit rule/norm invocation on communication behavior: do norms operate in a constitutive way, a regulative way, or both?
- What might explain the emergence of a neutral style of communication which does not appear to accord with the 'etiquette' and is contrary to the apparent reward for making positive utterances which are more likely to be validated (i.e. to be influential)?
- In Wikipedia, as elsewhere, norms have been observed to be violated in pursuit or defense of other norms: what does this reveal about factors influencing norm salience in on-line communities?

Activity theory furnishes a useful meta-analytical perspective. It provides a systemic view which focuses attention on complex interaction effects rather than assuming simple causality. What it does not do is provide guidance on how these effects may be further explored or explicitly tested. Some consideration is given to this here as a basis for future research. The hypotheses are presented here in approximately increasing order of the degree of cognitive capability and complexity demanded of the agents implied. The progression also represents an increasingly detailed account of the implied normative mechanisms. As a result further investigation of the hypotheses suggests the need for different methods. In the earliest examples there is a very significant scope for simulation to make a contribution as the cognitive complexity of agents is relatively minor. As the examples progress it is anticipated that the methods emphasis will need to shift towards more empirical work with perhaps a peripheral role for simulation or no feasible role at all given the current state of development of the technique.

Explaining the difference in quality

The following two hypotheses are suggested as a basis for explaining difference in the quality of articles. They are presented here as dichotomous although it is conceivable that elements of both may be operating at the same time. Within each there are several variations on the general theme.

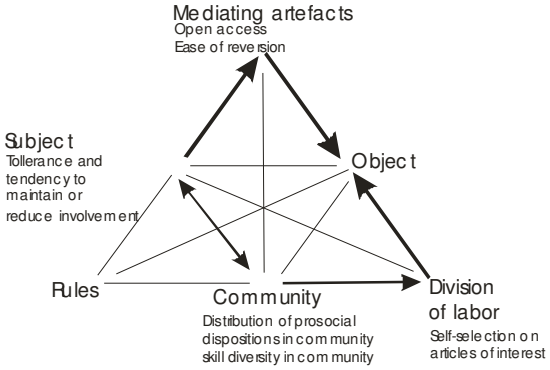
- It's all in the editing.
    - Chance association hypothesis
    - Primed attraction hypothesis
    - Wisdom of crowds hypothesis
- Norms influence discussions which influence article quality.

It is all in the editing

This hypothesis has several secondary propositions which are all based on the possibility that the dominance of one persons edits or of collective editing largely explains any difference in quality. This hypothesis places emphasis on subjects dispositions, motives and strategies, the effect of division of labor and the effect of the WikiWiki artifact and in particular the way in

which it supports open access and ease of reversion. This is represented diagrammatically as follows.

Figure two: Editing activity as determinant of article quality



Brandes et al (2008) have conducted analysis of editing behaviour in both Featured and Controversial articles. This analysis was performed on a random sample of 60 articles from each category and included an arbitrary selection (i.e. those which were not flagged as either Featured or Controversial). They found that both classes of flagged articles had a much higher number of revisions than the arbitrary. While the high number of revisions for Controversial articles (3059 compared to an overall average for Wikipedia of just 38) was expected it was less so for Featured (1382 revisions).

Both Featured and Controversial articles were also characterised by a higher than average bi-polarity index (one group of editors primarily changing the work of another group). They found that the bipolarity of Controversial articles was significantly higher than for Featured – an unsurprising result. The index of balance of authorship – a measure of the degree to which an identified groups contribute to the article – established that there was a greater balance of contribution between opposing groups in Controversial compared to Featured articles. This was interpreted by the authors as suggestive of greater attention being given to achieving a balance of perspectives in contested articles. Again bipolarity was found to be higher in both groups of flagged articles when compared to articles in general. A tentative explanation offered for this was that both tend to attract a disproportionate level of vandalism and that this accounts for an elevated level of bipolarity in both groups.

There are several sub-hypotheses which might explain these patterns.

Chance association hypothesis:

This hypothesis places primary emphasis on the compositional characteristics of the community from which subjects are drawn and on the character of individual editors. Regularity is explained by pre-existing social (normative) dispositions brought by individuals from outside. The idea that individuals display differences in pro-sociality, such as agreeableness and empathy, which are stable over time has long been accepted (Penner et al., 2005a). Chance variations in the pro-sociality of individuals drawn to edit a common article at a particular point in time could therefore explain differences in quality of the object. Change in the number of more pro-social to less pro-social individuals over time may lead to a change in status of the article. Featured sites are those that at the present time have more pro-social contributors who have, through their collective edits, achieved neutrality and coherence without a need to discuss it or to attempt to regulate one another. There is a contribution here from the characteristics of the WikiWiki platform. Firstly, it is the complete openness which supports access by a wide population of editors. Secondly, the ease by which vandalism can

be corrected using the reversion facility stacks the odds in favour of the pro-social over the anti-social.

The volunteer nature of Wikipedia suggests that levels of pro-social dispositions of editors should be higher than the population in general. This is due to their willingness to give time freely and due to a tendency for the most pro-social to persist for longer even though the burden of monitoring, policing and correcting as well as contributing is significant. While this may suggest that the community of Wikipedians has more pro-social character on average than the wider community it does not diminish the possibility that there is a distribution around this mean. This does not preclude the operation of norms and rules, indeed it could be argued that such individuals would likely be more sensitive to local rules and more willing to comply. However, it may also be the case that Wikipedia does not constitute a norm salient environment for many participants, pro-social or otherwise. The anonymity and resulting impersonality may contribute to a task focus regulated almost exclusively by individuals being guided by existing personal dispositions.

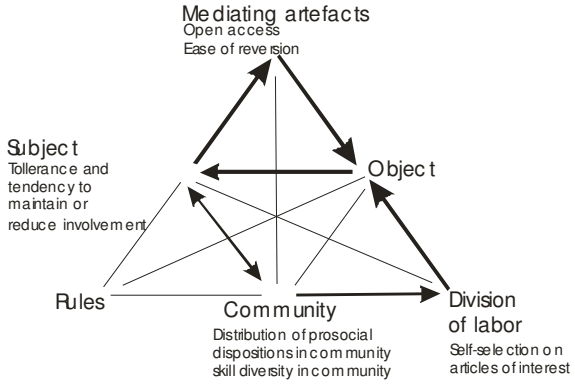
This hypothesis is the most parsimonious of all of those presented here. As such it should be actively examined before more complex explanations are sought. Some tentative support for this it is found in the multilevel analysis undertaken by Goldspink (2009) which shows that most of the variance between articles discussion styles is attributable to the individual and not to the article subject or topic of conversation. We also know that demographic characteristics and distributions matter as Pfeil et al (2006) have identified a relationship between editing behaviour and cultural differences measured using Hofstede's dimensions. This hypothesis could be readily tested by means of a simulation which assigns agents with different styles to a collaborative task and where the successful achievement of that task is dependent on the effectiveness of collaboration. Alternative runs which randomly assign agents to the task could ascertain if those tasks which happened to have allocated to them a disproportionate number of pro-social individuals were more effectively completed.

There are several variations on this theme which could be similarly explored.

Variant one: Primed attraction hypothesis:

The hypothesis outlined above assumes chance association with any given article. In Wikipedia editors are not randomly assigned but rather they self-select articles. It seems reasonable that some attributes of the article (such as subject or style) may influence who is attracted.

Figure three: Editing activity as determinant of article quality with primed attraction

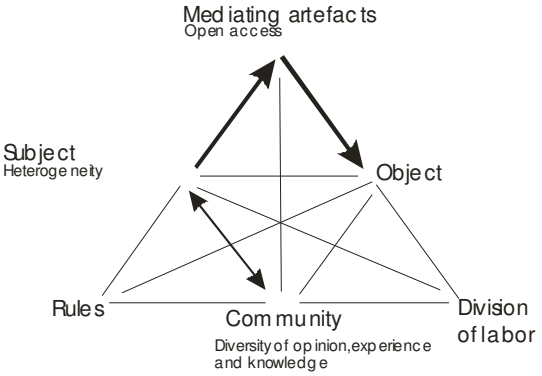


This hypothesis places the primary emphasis on the subject but sees a role for the existing object. The tendency for more pro-social rather than less pro-social contributors to be attracted to a particular article may be influenced by the initial quality attributes of the article. Biased articles may attract less pro-social protagonists for example. This process may also become self-reinforcing as the quality improves and/or the article is flagged as of high quality it may attract further pro-social individuals who amplify its quality attributes further. A test for this may comprise a simulation, similar to that outlined above, but where agents with different pro-social dispositions are attracted differentially to different article attributes.

Variant 2: Wisdom of crowds Hypothesis:

Surowiecki (2004) argues that given sufficient diversity and independence of opinion, decentralization of knowledge and a means to aggregate or bring these diverse opinions together, crowds can out-perform experts in many areas of prediction and problem solving. Through the lens of activity theory, this proposition places emphasis primarily on the characteristics and organization of community, diversity (heterogeneity) of subjects and the availability of suitable technical artefacts for aggregation of the diverse perspectives.

Figure four: Editing activity as determinant of article quality: Wisdom of Crowds



In contrast to the above hypothesis the focus here is on the distribution of diverse ideas rather than pro-social dispositions. Applied to Wikipedia the argument would be that given enough time and given that there are sufficiently diverse opinions brought to any article, the mean opinion will be a good approximation of the widely accepted (neutral) position. Wikipedia would appear to satisfy all of the required conditions but the existing evidence is mixed in its support.

In his exploration of quality in Wikipedia Andrew Lih (2004) proposed two proxy indicators. The first of these is **rigor** measured by the total number of edits undertaken on an article. The assumption was that more edits implied a deeper treatment of the topic and greater attention to its presentation and readability. The second measure was **diversity**, measured by the total number of unique editors. More editors were assumed to imply a wider range of perspectives contributing to depth and breadth of coverage and hence quality in a manner similar to that proposed by the Wisdom of Crowds hypothesis. These assumptions were not, however, backed by qualitative evaluation of the articles undertaken by Lih himself. Similarly Stvilia et al (2005) found that a distinguishing difference between Featured articles and articles chosen at random is that with the former there generally exists *'…a small core group of editors which is relatively homogeneous in terms of sharing social norms of cooperation, including communication protocols.'* (2005: 13) a finding directly at odds with the Wisdom thesis.

On the other hand Stvilia et al did find that Featured articles were on average 18 times larger than randomly selected articles and on average 3 times older (1153 days compared to 385) and thus subject to many more edits (257 compared to 8). The relationship between quality and time was also tested by Wilkinson and Huberman (2007: 11) who conclude that *'…a small number of articles, corresponding to topics of high relevance or visibility, accrete a disproportionately large number of edits'* and that *'Wikipedia article quality continues to increase, on average, as the number of collaborators and the number of edits increases'*. This suggests that there is a relationship between quality and the characteristics suggested by the Wisdom of Crowds hypothesis but it is linked to visibility: only highly 'visible' articles attract the attention of sufficiently diverse a population of editors for the effect to be realized. This is supported by research by Braendle (2005) whose analysis of 450 articles in the German Wikipedia indicated that **interest** (measured using number of edits, unique authors, traffic, age and number of links) and **relevance** (measured by the number of results which occurred in Google) are related to quality. The importance of visibility was supported by Lih (2004) who found an increase in quality after an entry had been cited in the mainstream press. The implication of these findings is that articles which attract little interest may persist for extended periods at a low state of development and may therefore be of poorer quality. This problem is partly addressed within Wikipedia by the capacity for editors to flag articles as 'NPOV', 'disputed' or 'accuracy disputed' etc. thereby focusing greater attention on them. The conclusion seems to be that for most articles there is insufficient diversity available due to self-selection and the resulting division of labour across the whole of the Wikipedia. An exception occurs when an article attracts significant attention due perhaps to media attention and for a time attracts sufficient diversity to undergo a marked shift in quality. Failing this, the dedicated attention of a few is more likely to improve quality.

So far we have made no account of how norms operate within the Wikipedia itself but rather assumed that there may be some distribution of norms within the wider community from which editors are drawn. It is, however, reasonable to speculate that as individuals interact in Wikipedia new norms may emerge, or that innovations may be needed to modify or reinvent protocols to guide the unique activity that this environment presents. Investigating this may also help us focus on the actual mechanisms of norms, which as Conte et al have noted, is still insufficiently explored (Conte et al., 2007).

What are the mechanisms of norms?

Before we consider hypotheses about normative behaviour proper in Wikipedia some consideration needs to be given to how it is envisaged norms 'work' based on existing social theory. The theory in this area forms part of an ongoing debate about a fundamental problem in social science – known variously as the problem of structure and agency or the micro-macro problem (Fuchs & Hofkirchner, 2005; Goldspink & Kay, 2004). This issue concerns the nature of the relationship between macro social structures such as institutions and norms, and micro processes such as the cognition of the human individuals that constitute that society, populate the institutions and enact (or not) the norms.

Gibbs argues *'Sociologists use few technical terms more than norms and the notion of norms looms large in their attempt to answer a perennial question: How is social order possible?'* (Gibbs, 1981). Not surprisingly then the concept has been incorporated into a wide range of alternative and often competing theories. All are concerned to explain social (more specifically behavioural) order but alternatively posit norms as a) a structural constraint on human agency or b) as a product of collective agency. A handful of dialectical theories posit them as both but struggle to specify the mechanisms and as a consequence these two perspectives are seldom brought together and even more seldom is the explanatory pathway made explicit. The failure to explain the generative mechanism being invoked has led to the suggestion that 'norm' is a generic concept (Gibbs, 1965). This suggests that it has no

explanatory value in and of itself - any invocation of the term needs to be explained and more precisely defined in the context in which it is being used. Examining the attempts to define the term leads to several variations:

> n sociology, a **norm**, or **social norm**, is a rule that is socially enforced. Social sanctioning is what distinguishes norms from other cultural products or social constructions such as meaning and values. Norms and normlessness are thought to affect a wide variety of human behavior. Source: Wikipedia

> *Norm also called* ***Social Norm,*** rule or standard of behaviour shared by members of a social group. Norms may be internalized—*i.e.,* incorporated within the individual so that there is conformity without external rewards or punishments, or they may be enforced by positive or negative sanctions from without. Source: Britannica

Gibbs (1981: 7) pulls these together with the statement *'A norm is a belief shared to some extent by members of a social unit as to what conduct* ***ought to be*** *in particular situations or circumstances.'* The missing explanation pertains primarily to how it is that a) beliefs come to be shared b) a common deontic is applied to c) situations evaluated in broadly similar ways by different social actors. Also missing is any argument as to how a norm acquires a particular valence (good or bad) associated with particular contexts. This process of recognition and attribution implies some sort of cognitive mechanism. Within cognitive theory there are two alternative paradigms cognitivism and enactivism and the mechanism suggested by each are very different.

Cognizing norms

Within cognitivism, the emphasis is placed on processes of reasoning (Hayles, 1999). Cognition is cast as a mechanism whereby agents construct maps (internal representations) of objective worlds and reason about their relationship to that world. All rational choice, game theoretical approaches, as well as Beliefs Desires and Intentions (BDI) approaches are of this paradigm. Cognitivism accords with folk views of intelligence – the mind as computer– and is the approach typical of first generation AI (Franklin, 1998). Reason is taken to be function of a disembodied mind, perhaps subject to biological (and evolutionary) constraint but nevertheless largely independent of it. Research undertaken within the EMIL project clearly indicate that past approaches to simulating norms can be shown to fall exclusively into this school of thought. There is, however, an alternative.

Within an enactivist view of cognition much greater emphasis is placed on the role of affect. The work of Antonio Damasio (2006) is relevant here – Damasio argues from the perspective of embodied neurophysiology that we use somatic markers to attribute valence to situations and that while these apply at the level of the individual they may become more shared. These markers play an important pre-filtering role in decision making. Others have also argued for a significant role of affect in social interaction, including for the establishment of moral norms. Evan Thompson, for example, quotes Frans de Wall as stating *'Aid to others in need would never be internalized as a duty without the fellow-feeling [sympathy] that drives people to take an interest in one another. Moral sentiments come first, moral principles second…'* (Thompson, 2001). Thompson's own argument is that empathy is an innate biological capability that is fundamental to the possibility for our understanding of self – that is, of being self-conscious. It provides the foundation for all sociality enabling us to be open to and led by others experience and thus shaping all of our social interactions. Empathy is an embodied capacity – which is to say that it operates through the somatic as well as neural and sensori-motor system. From the enactive perspective intelligence moves from problem solving capacity to flexibility to enter into and engage with a co-constructed world. In this context norms are markers of this shared character but they are not 'in' agents brains or 'in'

environmental 'structures' (although they may be reflected in artefacts), rather they are recreated instant by instant in the reciprocal interactions of agents based on those agents unique ontogonies which are themselves a product of the agents histories, at least part of which may have been shared with other agents (De Jaegher & Di Paolo, 2007). The enactive worldview shows how we are continually creating worlds as we interpret them. Our interpretations predispose us to act in a certain way and this comes together to 'unfold' a reality or *'lay down a path in walking'* (Varela et al. 1992). This approach shifts the conception of norms away from the reified view which casts them as stable cultural meaning structures with a denotative character and implicit deontic. A norm invocation may be signalled behaviourally or linguistically but it represents not so much a command as a perturbation by one agent in order to shift the *umwelt* of others or at least to change the interaction dynamics. This may or may not be done consciously. This approach is attracting considerable attention in evolutionary robotics (De Jaegher & Di Paolo, 2007; Di Paolo et al., 2007), artificial life (Barandiaran, 2005), and distributed theories of language (Cowley & Macdorman, 2006). So far it has had little direct application to norms and norm simulation. It is, however, reflected in some considerations of the effect of computer mediated communication. Riva and Galimberti (1998: 434) for example take the position that communication is *'…not only – or not so much – a transfer of information, but also the activation of a psychosocial relationship, the process by which interlocutors co-construct an area of reality'*. Where communication is computer mediated *'the co-presence of utterances, rather than the physical co-presence of interlocutors, is now seen as the key to the construction and performance of cognitive functions'* Here they focus on the centrality of micro coordination through communicative exchange, citing Goffman and his concept of the interaction order (1983). The connection to enactive cognition is reinforced when Riva and Galimberti state *'Thus, cognition is now seen as something that happens between rather than inside subjects; as a media-structured loop that begins and ends with the subjects themselves; as a continuous exchange which generates a shared construction of reality as the interface between the individual and collective; as cognition and interaction, mental activity and social activity.'* (Riva & Galimberti, 1998).

Emergence and immergence

While the use of the concept of emergence is longstanding it is nevertheless controversial. Emergence is associated with what in contemporary terms are called complex or self-organizing systems. There have been several important contributions to the debate in recent times where it has been observed that the type of emergence which occurs with simple physical particles will be different from that which emerges between intelligent agents. Gilbert, for example distinguishes between first and second order emergence. First order emergence includes macro structures which arise from local interactions between agents of limited cognitive range (particles, fluids, reflex action). By contrast, second order emergence is argued to arise *'where agents recognize emergent phenomena, such as societies, clubs, formal organizations, institutions, localities and so on where the fact that you are a member or a non-member, changes the rules of interaction between you and other agents.'* (Gilbert, 2002). This reflects higher order cognition on the part of the agent. In particular it reflects a range of capabilities including, but not limited to, the ability to distinguish class characteristics; assess 'self' for conformity with class characteristics and/or signals from other agents which suggest acceptance or belonging; and the ability to change rule associations and behavior as a function of these changes. First and second order emergence then each imply qualitatively distinct mechanisms and suggest a continuum of orders of emergence linked to cognitive capability.

This was Castelfranchi's observation also when he coined the term immergence as a complement to emergence. Castelfranchi (1998b: 27) refers to this as cognitive emergence and states *'Cognitive emergence occurs where agents become aware, through a given 'conceptualization' of a certain 'objective' pre-cognitive (unknown and non deliberated)*

phenomenon that is influencing their results and outcomes, and then, indirectly, their actions.' Thus Castelfranchi conceives of a feedback path from macro pattern to micro behavior in much the same way as Gilbert, except that here a cognitive mechanism is specified and in this case it is a associated with the cognitivist school of thought. Castelfranchi argues that this mechanism has a significant effect on emergence and indeed *'characterises the theory of social dynamics'* – that is, it gives rise to a distinct class of emergent phenomena. In this account, the representations agents have about the beliefs, desires and intentions of other agents plays a causal role in their subsequent behavior and therefore shapes the structures they participate in generating.

These ideas are more comprehensively reflected in the five orders of emergence suggested by Ellis (2006:99-101). These are:

1. Bottom up leading to higher level generic properties (examples include the properties of gases, liquids and solids)
2. Bottom up action plus boundary conditions leading to higher level structures (e.g. convection cells, sand piles, cellular automata)
3. Bottom up action leading to feedback and control at various levels leading to meaningful top down action - teleonomy (e.g. living cells, multi-cellular organisms with 'instinctive' – phylogenetically determined reactive capability)
4. as per 3 but with the addition of explicit goals related to memory, influence by specific events in the individuals history (i.e. learning)
5. In addition to 4 some goals are explicitly expressed in language (humans).

Ellis's framework makes clear that the range and type of emergence possible in a system depends fundamentally on the range and class of behavior that agents are able to generate and that this varies depending on the properties of the agent. Goldspink and Kay (2007; 2008a; 2008b) have further developed on these themes and argue that these orders may be present at the same time and the effects will interact generating even more complex emergent outcomes.

The hypotheses pertaining to the origin of order and quality in Wikipedia considered so far have focused on stochastic effects with or without the presence of simple reinforcing feedback loop and/or simple learning.: there has been no need to invoke more sophisticated cognitive capability. As we consider the possible effects of norm innovation in Wikipedia there is a need to consider how norms may be taken to operate based on the alternative conceptions of human cognition considered above. While the directive hypothesis of norms commonly embraces a cognitive position which assumes that agents receive an implicit or explicit command, recognize it as such, evaluate it against some criteria of relevance and then choose whether or not to act on it as reflected in the EMIL-A architecture proposed by Conte et al (Andrighetto et al., 2007; Conte et al., 2007), the hypotheses set out below support several alternative cognitive mechanisms, including an enactivist one.

In the cognitivist approach suggested by Conte et al, norm recognition involves the transmission of deontic commands between members of a community or by artifacts. If an input so transmitted is recognized as a norm it becomes a *belief* in the mind of the recipient to the effect that *"there exists a norm prohibiting, prescribing permitting…x"* This belief is then checked for pertinence by means of a lookup table which also specifies goals appropriate to the norm based on an assessed degree of salience. This salience may be objective – taken to apply to all agents; or subjective – based on the individual history and experience of an agent. Inputs which do not correspond to an existing norm for the agent have to be evaluated to ascertain if they satisfy the criteria of a norm.

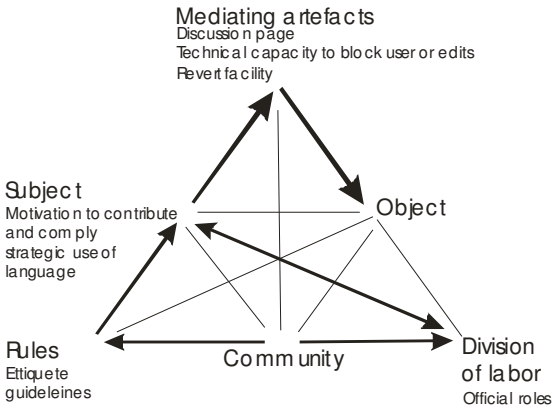
From an enactive perspective, the recognition involves the complex presenting situation triggering a somatic marker so as to change the editors future disposition to action. This affective filtering uses past experience to influence subsequent action at both unconscious and

conscious levels. This approach can therefore embrace situations where norm following appears automatic and happens out of awareness (without reducing this to simple mimicry) as well as norm following which results from rational deliberative action.

Norms influence discussions which influence article quality

When we began the Wikipedia case research associated with EMIL, this was our starting hypothesis. It assumed that quality would be more likely where dialogue dominated discussion and where violations of protocol or of stylistic conventions were regulated by the explicit invocation of rules and norms and/or by the use of illocutionary force of language (Searle, 1969) to achieve compliance on the part of violators. Presented diagrammatically the relationships implied are as shown in figure two.

Figure five: Mechanisms of Normative Influence



From an activity theory perspective the attainment of an object of high quality was assumed to be achieved by the voluntary compliance of subjects with rules and, should this fail, by the explicit invocation of rules by others subjects supported by norm innovation – the creation of new norms and rules appropriate to an unmet regulatory need. The WikiWiki platform facilitated this through the provision of a forum for discussion, made it simple to correct the contributions of non-compliers through the reversion facility and in extremis, supported the use of formal sanctions through the community based election of 'officials' with technical control over members access to articles.

This hypothesis was partly explored (Goldspink 2007, 2008) by our detailed analysis of utterances on Discussion pages of a sample of Featured and Controversial articles. It was concluded that the detailed and specific behavioural etiquette published in Wikipedia had little apparent influence on the overall character and style of communication on Discussion pages. For example, while Reagle (n.d) argues that the practices advocated by various Wikipedia policies *'…are very much aligned with Yankelovich's (2001) notion of dialogue which relies upon the notion of empathy'*, on Discussion pages, the markers of quality dialogue were generally absent. There was, for example, little evidence of perspective talking, or seeking to understand the view point of others through paraphrasing, reflecting understanding and questioning.

It was noted also that the invocation of rules or norms was rare. This also suggests that the rules may have limited impact on actual behaviour. There was a lack of evidence of active negotiation of expectations and standards and convergence of behaviour towards a norm. While there was a difference between Controversial and Featured article Discussion pages this was small and the quality of the interaction could not explain the difference in status. Similarly there appeared to be little in the subject matter of the two groups of articles that would explain the difference – both contained subject matter which was contestable and subject to significantly diverse opinion. Multilevel analysis of this data suggests that

differences in style of communication are primarily attributable to the individual rather than the topic of discussion or the subject of the article.

On the face of it this evidence does not support the regulatory operation of norms and rules although there is some uncertainty about whether the invocation of rules operates in a longer term way or in a diffused way as is discussed in a later section. It is feasible that knowledge of the rules leads to a level of voluntary compliance. There was some evidence that illocutionary force may be a factor although the high incidence of ignoring of directive utterances, particularly by 'out-group' members suggests that this is limited and may be based on group identity. While normative influence may not work in an entirely directive way it is possible that more discreet or subtle mechanisms are at work. We can now examine some of these in more detail.

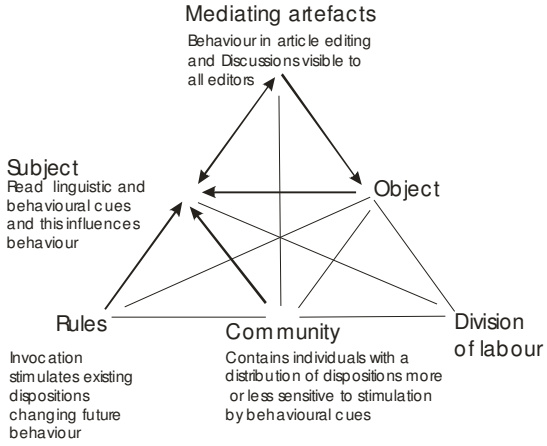
The lack of apparent effect of explicit norm invocation on communication behavior.

Three hypotheses are explored here, again each suggests several variants.

- Diffused effect hypothesis
- Group salience hypothesis

Diffused effect hypothesis

Figure six: Indirect and diffused normative influence



While this hypothesis assumes that norms do work by a directive mechanism – at least in part – it allows for several alternative pathways based on alternative cognitive models.

Taking the EMIL-A architecture as a state of the art example of a cognitivist approach, norm communication can happen by means of tacit or explicit signalling between agents. This may happen by their edit actions, as well as by communicative acts on Discussion Pages and or Talk pages. In addition, rules are explicitly communicated on pages which identify rules of etiquette and guidelines. In the context of Wikipedia, this signalling takes place in public spaces and is therefore available to all who access those spaces – there is no possibility for one on one targeting of communication of norms. In practice the Talk pages are probably less frequently visited, providing the Wikipedia equivalent of the office 'water cooler' for more discreet directed discussions, but still with some chance of being 'overheard'. The analysis of Discussion pages undertaken by Goldspink (2008) suggests that this openness does influence the way in which editors attempt to influence one another on the relatively rare occasions that they do explicitly invoke a rule or norm. This evidence allows that the signalling may be subtle and associated with a wider range of behaviours and artefacts, including the style of communication in Discussion and Talk pages and the actual editing behaviour. It also allows

for other potential influences on the part of the subject, including the reading of cues to do with group affiliation. Goldspink's research showed that norms and rule invocations observed in the Discussion pages were often undirected – made as a general statement to nobody in particular such as ' this is not the way we do things in Wikipedia'. Consequently, while the evidence suggests that the most common immediate response to an invocation may be to ignore it, the effect may more diffused than is commonly assumed in the directive approach to norms. Rather than having an immediate effect on a target who has to interpreted the deontic command and decide whether to comply or reject the invocation, it may effect a specific target and/or non specific targets alike. This may change their likely future behaviour including their likelihood of engaging in non normative behaviour. While this is not inconsistent with the norm recognition approach suggested in EMIL-A (except that all messaging is broadcast rather than directed) it suggests that norm recognition within this community acts on prior socialization of the participants – that is to say it reinforces the importance of subjective (unique individual) salience based on each agents prior history and current participation in a number of different (intersecting) social domains. Here explicit as well as tacit invocations may serve to reinforce previous socialized norms or to remind participants of constitutive rules or to shift other dispositions and goals. Norms and rules do not have to operate as primary controls on behavior but exercise influence by acting on the minds of editors in several ways. These editors are participating in many social domains of discourse and action which can be expected to both effect and be effected by what happens in their participation in Wikipedia. Norms within different domains may differ or even contradict and the agent confronts a dilemma about which to follow. If Wikipedia is regarded as just one form of media used by individuals as they go about their lives then it may be expected that events in the virtual space will be interpreted by them in this wider social context and in accordance with alternative norms which prevail there.

The manifestation then is of a diffused effect both across participating individuals and potentially also over time as well as across the boundaries of various social domains in which editors participate. Linguistic interventions in the form of explicit norm or rule invocation or the use of illocutionary force in linguistic utterances, as well as behavioural interventions such as corrective editing and reversion may have a delayed and diffused effect or no effect at all on non-compliant individuals depending on their wider context and prior history of interactions in Wikipedia and other social domains. It is not immediately clear how this may be simulated as typical computational agents (particularly in the cognitivist tradition) lack a 'social history' and generally exist in a single social domain.

Group salience hypothesis

The area of social research which has arguably most attempted to come to terms with the problem of social boundaries and its effect on norms is social identity theory. Within social identity theory, norms have been defined as '*…the emergent phenomena of human association'* (Hogg & Abrams, 1988) – strongly suggesting a constitutive rather than a directive orientation. Unlike sociology which is concerned with the effect of norms across wide social aggregates, identity theory is concerned with the relationship between norms and groups. Terry et al (1996; 1999) have argued that norms will impact on the causal relationship assumed to operate between attitude and behaviour to the extent that they are perceived as salient within a reference group important to the actor, in particular, a group which is important as a basis for the self-identity of that actor. In other words their research has demonstrated a strong link between the effect of a norm and the source of that norm.

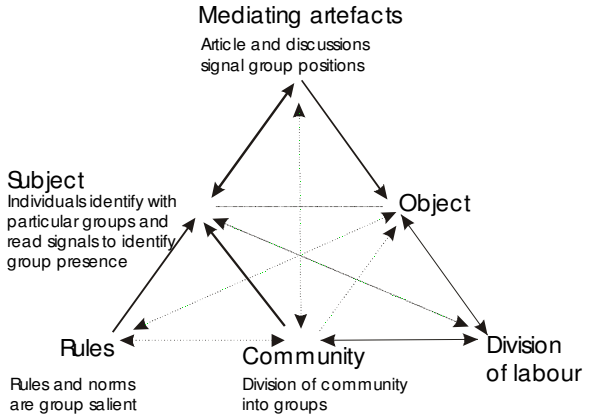
> *'…high identifiers are intrinsically motivated to engage in group mediated behaviour and to conform to in-group norms. In contrast, low identifiers exert themselves on behalf of the group only in the presence of external pressures, such as accountability, that affect their self-presentational concerns and their desire to achieve positive evaluation by others'. (Smith et al., 2007: 241)*

Rules and norms salient within a group with which the subject identifies, will influence the effect of that subjects pre-socialised attitude on their behaviour in particular contexts. Applied to Wikipedia the subject will use cues in various artefacts, including the article itself as well as Discussions and Talk pages to read the group environment and to identify other subjects with particular groups. Where the subject does not identify with a group, norms of that group will have little or no effect on his/her behaviour and the individual will act in a manner consistent with their individual attitudes: he/she will ignore invocations by members of a perceived 'out-group'.

As the title of the theory suggests identity plays a central role. *'group norms have most impact on individuals for whom the group is an important basis `for self identification'* (Smith et al., 2007: 240). Identity has also been argued to play a significant role in motivation to participate in volunteer projects such as Wikipedia (Hemetsberger & Pieters, 2001; Rullani, 2005). The importance of identity presents some distinctive challenges for both theorising and modelling the effect of norms as it implies reflexivity or self-reference and associated orders of emergence (see Goldspink & Kay, 2008b for an examination of the implications of this and a discussion of normative mechanisms).

Cialdine et al (1991) and Kallgren et al (2000) also hypothesised and subsequently demonstrated through a series of experiments that the effect of a norm on an individuals behaviour was influenced by the focus (salience) of the norm and the degree to which a norm was internalised – i.e. held as personally significant. Personal significance points again to issues of personal history, importance of current context and agents participation in intersecting social domains. These perspectives place a significant emphasis on the structure of community and the nature and history of subject.

Figure seven: Social Identity and normative action in Wikipedia



Postmes et al (2001: 1243) has considered the effect of anonymity on group related norm compliance. They note that *'identifiably, as opposed to anonymity, would be expected to enhance immediacy and social presence and thus facilitate social influence.*' However, imagined or implied presence may have an influence and hence it may not be the case that immediate presence is necessary for influence. This is supported by self-categorisation theory, which has it that social influence is *'in the first instance cognitively mediated by ones self-categorization as a group member, rather than by processes involving social contact per se.*' accordingly *'the strength of group self-categorization is closely bound up with the affective and emotional significance attached to this self-definition…'* (Postmes et al., 2001: 1244) This reinforces the possible role of affect and reflexive identity in the mechanism of norm recognition and response and these are mechanisms not generally included in cognitive approaches to the study of normative mechanisms. In further considering the effect of computer mediation of interaction in Wikipedia the SIDE (Lea et al., 2001) model suggests that when a social identity is already salient, visual anonymity can enhance group salience

due to its diminution of obvious individual difference and hence interpersonal concerns in favour of group characteristics. Recognition of salience is a non-trivial task and cognitively suggests analogical reasoning or pattern recognition, possibly involving automatic affect or somatic marker possibly followed by a capacity to recognise (conceptualise) a situation and fit it to a set of dimensions and to associate those conceptual dimensions with past experience to which has been associated a norm. In terms of the cognitive processes involved, Bicchieri argues that the recognition process is automatic and unintentional (i.e. not the result of rational deliberation). In other words the response is due to the person recognising an existing norm as salient and generating a response (initiating a behavioural script) in response to cues.

The effect of group identity in a wiki editing environment is being examined empirically by means of an experiment being run at Koblenz University.

The emergence of a neutral style of communication which does not appear to accord with the 'etiquette'.

Earlier findings by Goldspink point to the emergence of a form of discourse in the Discussion pages which is distinctive and difficult to explain given the etiquette of Wikipedia, the nature of the task environment and given that positive style is rewarded and negative style punished. This appears to be a genuinely emergent communication style norm in Wikipedia Discussion pages. The following are among the hypotheses which may in part or in various combinations suggest some explanation for this.

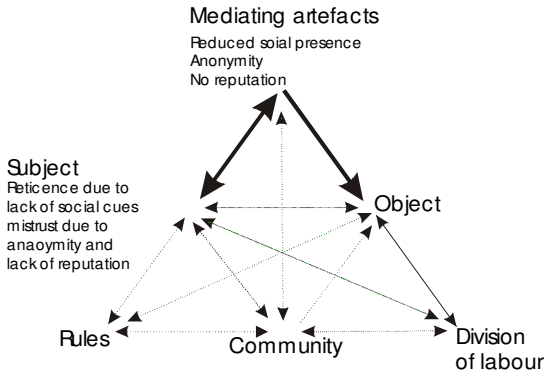
- Suspicion hypothesis
- Cue matching (copying of convention)
- ersatz academic speak – how non-academics think academics converse.
- Type of person attracted
- Pareto sub-optimal hypothesis (complex effect)

Suspicion or Uncertainty hypothesis

One of the well studied social effects of computer mediation is with the way in which it influences social presence and how changed cues of social presence influence communication behaviors (Biocca et al., 2003; Biocca et al., 2002). The absence of social cues has been argued to support greater social reserve as well as greater social volatility – a tendency to 'flame' or be abusive. One of the effects of reduced social presence is its impact on trust. This is further exacerbated in Wikipedia by anonymity and the associated ambiguity about the individuals goals, knowledge base, group identity, credibility and motives. Following Goffman (1983), uncertainty has been argued to reduce trust. Conversely normality – that is stability and predictability, including evidence of norm compliance – has been argued to be an important contributor to the creation of the possibility for trust (Misztal, 2001). Rules and norms, when followed, make the behavior of others more predictable reducing uncertainty and the attendant anxiety. The observed adoption of a neutral tone in Wikipedia Discussions may serve two purposes a) to reflect the 'arms length' and indifference that signals low trust b) to reduce the likelihood of triggering adverse behavior in others. The stability of this emergent pattern is, in the context of Wikipedia Discussions, a form of normality. As an article becomes polarized a neutral impersonal style reduces what individuals reveal about themselves and about their intentions while nevertheless becoming more predictable in adopting a style which is non-offensive and discouraging of effusive interaction which could trigger more emotive interactions by others. *'Acting normally, achieved by making a self-conscious effort to follow interactional rituals, affirms the collective image of what is a normal manner of conduct and in turn ensures trust and tacit cooperation.'* (Misztal, 2001: 316). The social presence position is compatible with this view as it is argued that social cues play an important role in signaling and the reading of behavioral and emotional signals leads

to a tacit knowledge *'…that enables them to understand the meaning of action within a particular environment.'* (Misztal, 2001: 314).

Figure eight: Managed communication in an environment not supportive of high trust

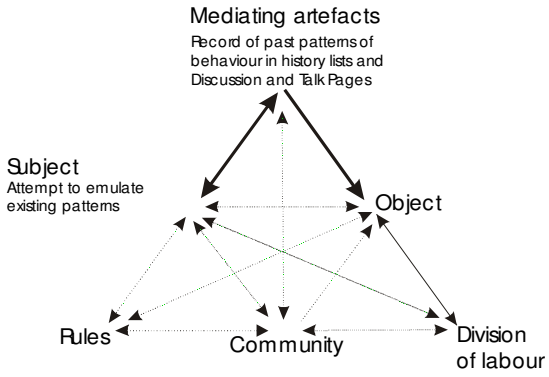


This suggests that the technical artifacts and individual subjects are a major determinant of the social dynamics. The artifacts fail to support exchanges which contain sufficient social information to support more intimate communication and to foster confidence through reputation.

Cue matching (copying of convention)

The pattern may also be explained as a product cue matching or social priming. This hypothesis argues that newcomers to a Discussion or to the editing of an article are likely to adopt behavioral patterns similar to those that already exist – in essence taking these patterns as cues about the norms prevailing in that social domain. This is consistent with norm following being seen as an instance of copying or imitation. This imitation may work in a positive way (as with editors noticing the genre specific character or characteristics of quality in a Wikipedia article and seeking to copy it and hence potentially extending and amplifying those qualities). Alternatively it may have a negative effect where pre-existing negative style of communication in a Discussion page lead to further instances or an amplification of the negative style. What is missing here is how the observed neutral style became established in order to be copied.

Figure nine: Subjects copy artefactual cues indicating communication norms



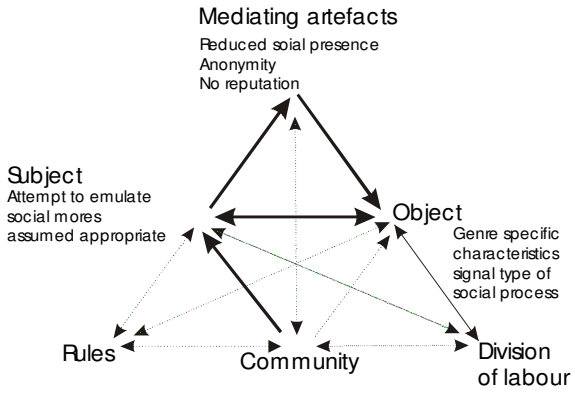
Clearly if this hypothesis is to be taken seriously then more would need to be known about how subjects choose what to attend to (selective attention and priming theory) and how individuals with different backgrounds and dispositions may attend and imitate different patterns as well as how the availability of artifacts that embed patterns in edits, discussions and talk interact to shape behavior in different aspects of the task. The principle cognitive process to be dealt with is that of pattern recognition, although subsequent decision making about which to follow could suggest rational and affect based decision mechanisms.

The effect of the priming or cuing effect of prior discussions in a wiki editing environment is being examined empirically by means of an experiment being run at Koblenz University.

Ersatz academic speak – how non-academics think academics converse.

This is a derivative of the cue matching hypothesis outlined above, except that the patterns being attended to are associated with the object itself (in this case a credible encyclopedia which conforms to the generally accepted genre characteristics of a repository of established facts) as well as patterns of community practice commonly associated with such objects. Encyclopedias are generally assembled by qualified experts – academics and specialists who, it may be assumed, have a particular way of working including a way of structuring discourse. It has been suggested that Wikipedia represents a rare opportunity for lay people to make a contribution to an information source of this type – in effect to become authors of a widely used public product. In taking up this challenge they may attempt to emulate what they assume is the style appropriate to the task, but, being unfamiliar with that style, replicate it poorly.

Figure ten: Replicating (poorly) behavior of subjects normally associated with a task of this type



For many non-academic editors, the Wikipedia may present a very rare opportunity to contribute to a written artifact which will likely persist for some time in the public domain. Wikipedia is different to many other Web II opportunities to participate in this type of productive activity, such as blogs, in that it requires a level of rigor more typical of academic endeavor. This includes the need to strike a neutral style, to balance alternative perspectives, to consider alternative evidence and to reference material – linking it to a wider body of credible literature. One possibility then is that they style of communication entered into on Discussion pages is what many who have no direct experience of academic discourse imagine it would be done. The lack of demographic information and/or any firsthand accounts from editors about the stylistic influences makes it difficult to test this hypothesis.

Type of person attracted

Other than the speculation on the level of pro-sociality of Wikipedians and the obvious fact that they need to be computer literate, little is known about the demographics of Wikipedia editors. Many of the questions raised could be advanced if more were known about this. Given the sheer scale of the exercise and the diversity of subject matter it is reasonably safe to assume that this community, taken as a whole, is quite diverse with respect to age, educational background, ethnicity, first language and professional background. It might, therefore, seem unlikely that any regularity in editing or Discussion behavior might result from an underlying similarity in the type of person attracted to become an editor. However Pfeil et als (2006) finding of cultural differences between different language Wikipedia suggest that, at a very broad level at least, this can occur. Similarly, to the extent that certain sorts of people are drawn to certain topics or by chance association as discussed earlier some patterns may emerge for short periods of time so it cannot be ruled out.

Pareto sub-optimal hypothesis (complex effect)

In economics the self-organized operation of free markets is argued to result in a Pareto optimum distribution of benefit. The linear systems assumptions which apply to the models which predict this outcome also suggest that there will be a single solution which represents such an optimum. However, within game theory, sub-optimum outcomes are shown to be possible (Axelrod, 1984; Axelrod & Keohane, 1985) and within complex systems theory the existence of rugged coupled fitness landscapes predicts that self-organizing systems often become stranded on sub-optimal fitness peaks (Kauffman, 1987, 1993; Kauffman & Macready, 1995). These are consistent also with the Satisficing solutions to complex problems identified by Cyert And March (1992) in Organization Theory. So, if the order observed in Wikipedia is the result of self-organization, driven by the complex (and non-linear) interplay of various aspects of the activity system, then it might well be expected that some of the emergent patterns appear to contradict what would be expected if rational action were assumed. In other words, at least some of the observed order would be contrary to what would be expected if agents rationally assessed the needs of the task and the relevance and salience of norms and acted accordingly.

This hypothesis suggests that some or all of the previously identified mechanisms may make a contribution to the emergent dynamics which we see in Wikipedia, including the generation of the distinctive pattern of discourse observed in the Discussion pages. It was noted above that the task of producing quality articles, particularly with respect to controversial topics, calls for quality dialogue yet the affordances of the technology remove some of the necessary qualities of communication which would support it. The anonymity and subsequent reduction in trust may come into tension with the goal needs and result in a compromise or 'satisficing' emergent solution – one which balances the competing forces and is difficult to explain from a rational expectations perspective. As a consequence, even though the evidence shows that negative style behavior is least influential, neutral moderately and positive most, the neutral style may be the pareto-sub-optimal outcome: a 'satisficing' solution that falls short of the optimum based on theoretical propositions as well as empirical observations of what might constitute effective practice for this type of undertaking. This hypothesis is essentially arguing that Wikipedia is a complex social domain which, while sharing many things in common with other social domains, nevertheless generates some patterns which are genuinely emergent and unique to it. This is not a very satisfying hypothesis although it is probably the most realistic. It signals that, as with all real world social systems, the complexity is such that it is difficult to gain traction analytically. We have a surprising result – consistent with genuine emergence – but struggle to explain it using available theory.

This should be the sort of hypothesis to which computer simulation is ideally suited. However, this assumes that we can model sufficiently well the critical characteristics of the agents. To the extent that this requires us to model more advanced cognitive capability, especially phenomena such as reflexive identity, then we are clearly a long way from being able to do so. Nevertheless simulation may enable at least some of the effects of different systems components to be identified.

Conclusions

There exist a range of alternative hypotheses about the operation and role of norms within online communities such as Wikipedia. The existing empirical evidence is patchy and more needs to be known in several critical areas, such as the individuals characteristics and motivation of individuals in order to support the empirical separation of competing explanations. Some of the hypotheses do lend themselves to further testing through simulation and simulation can also offer an alternative means – allowing experimentation as a partial substitute for empirical evidence in determining which hypotheses might best be supported. However there are limits to what is possible using current state of the art technology. Paradigmatic choices also place constraints on the range and type of behaviors which can be

explored by simulation methods. Cognitivist approaches to simulation lend themselves to the exploration of determinative approaches to normative action while constitutive normative mechanisms, particularly those which imply some role for affect, may be better pursued using alternative conceptual and simulation methods such as those currently being developed in association with enactive views of cognition.

Past attempts to model norms have drawn on a very limited set of social theoretical roots and have adopted very limited and simplistic assumptions about both the nature of social interaction and cognition. This is of course normal – the aim of modeling is to capture the most parsimonious representation of mechanisms which explain a given phenomena. However, there are some social behaviors which simply cannot be generated without agents with considerable cognitive sophistication and in several cases (such as reflexive self awareness) we are a very long way from being able to incorporate even proxies of this capability into social simulations.

Several of the hypotheses set out here will be further explored using empirical methods as well as through the simulator. In EMIL we have designed an experiment with real subjects working on a Wikipedia type task which will allow us to further explore the pro-social dispositions aspects, primed emulation of communication style and the effect of group identity on communication in a wiki like environment.

References

Andrighetto, G., Campenni, M., Conte, R., & Paolucci, M. 2007. On the Immergence of Norms: A Normative Agent Architecture, Emergent Agents and Socialities: Social and Organizational Aspects of Intelligence Symposium. Washington.

Axelrod, R. 1984. The Evolution of Cooperation. New York: Basic Books.

Axelrod, R. & Keohane, R. O. 1985. Achieving Cooperation Under Anarchy: Strategies and Institutions. World Politics, 38: 226-254.

Barandiaran, X. 2005. Behavioral Adaptive Autonomy. A Milestone on the ALife roue to AI? San-sebastian, Spain: Department of Logic and Philosophy of Science, University of the Basque Country.

Becker, B. & Mark, G. 1997. Constructing Social Systems through Computer Mediated Communication. Sankt Augustin, Germany: German National Research Center for Information Technology.

Berger, P. L. & Luckman, T. 1972. The Social Construction of Reality: Penguin. ISBN: 0140600019

Bidwell, C. E. 1966. Values, Norms, and the integration of Complex Social Systems. The Sociological Quarterly, 7(2): 119-136.

Biocca, F., Harms, C., & Gregg, J.; The Networked Minds Measure of Social Presence: Pilot Test of the Factor Structure and Concurrent Validity, www.mindlab.org

Biocca, F., Harms, C., & Burgoon, J. K. 2003. Towards a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. Presence, 12(5): 24.

Braendle, A.; Many Cooks don't spoil the broth; Sept. 16, 2006, http://meta.wikimedia.org/wiki/Transwiki:Wikimania05/Paper-AB1

Brandes, U., Kenis, P., van Raaij, D., & Lerner, J. 2008. Modeling and Analyzing the Edit-Network among Authors

of Wikipedia.

Castelfranchi, C. 1998a. Through The Minds of the Agents. Journal of Artificial Societies and Social Simulation, 1(1).

Castelfranchi, C. 1998b. Simulating with Cognitive Agents: The Importance of cognitive emergence. In J. S. Sichman & R. Conte & N. Gilbert (Eds.), Multi-agent Systems and Agent Based SImulation. Berlin: Springer.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. 1991. A Focus Theory of Normative Conduct: A Theoretical refinement and reevaluation of the role of norms in human behavior. In L. Berkowitz (Ed.), Advances in Experimental social psychology: 201-234. San Diego: Academic Press.

Ciffolilli, A.; Phantom Authority, self-selective recruitement and retention of members in virtual communities: The case of Wikipedia; 12, 8, http://firstmonday.org/issues/issue8_12/ciffolilli/index.html

Conte, R., Andrighetto, G., Campenni, M., & Paolucci, M. 2007. Emergent and Immergent Effects in Complex Social Systems, Emergent Agents and Socialities: Social and Organizational Aspects of Intelligence Symposium. Washington.

Cowley, S. J. & Macdorman, K. F. 2006. What baboons, babies and Tetris players tell us about interaction: a biosocial view of norm-based social learning. Connection Science, 18(4): 363-378.

Cyert, R. & March, J. G. 1992. A Behavioral Theory of the Firm (2nd Edition edition ed.): Wiley Blackwell.

Damasio, A. 2006. Descartes Error. London: Vintage Books.

De Jaegher, H. & Di Paolo, E. A. 2007. Participatory Sense-making: An enactive approach to Social Cognition. Phenomenology and the Cognitive Sciences, forthcomming.

Demil, B. & Lecocq, X. 2003. Neither market or hierarchy or network: The emerging bazaar governance: 36: Université Lille/Institut d'Administration des Entreprises.

Di Paolo, E. A., Rohde, M., & De Jaegher, H. 2007. Horizons for The Enactive Mind: Values, Social Interaction and Play. In J. Stewart & O. Gapenne & E. A. Di Paolo (Eds.), Enaction: Towards a New Paradigm for Cognitive Science. Cambridge MA: MIT Press.

Donath, J. S. 1998. Identity and deception in the virtual community. In P. Kollock & M. Smith (Eds.), Communities in Cyberspace. London: Routledge ISBN 0415191408.

Ellis, G. F. R. 2006. On the Nature of Emergent Reality. In P. Clayton & P. Davies (Eds.), The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion. Oxford: Oxford University Press ISBN: 0199287147.

Engestrom, Y., Miettinen, R., & Punamaki, R.-L. 1999. Perspectives on Activity Theory. New York: Cambridge University Press.

Fjeld, M., Lauche, K., Bichsel, M., Voorhorst, F., Krueger, H., & Rauterberg, M. 2002. Physical and Virtual tools: Activity Theory Applied to the Design of Groupware. Computer Supported Cooperative Work, 11: 27.

Forte, A. & Bruckman, A. 2008. Scaling Consensus: Increasing Decentralization in Wikipedia Governance, 41st Hawaii Conferene on Systems Sciences. Hawaii: IEEE.

Forte, A. & Bruckman, A. n.d. Why do people write for wikipedia? Incentives to contribute to open-content publishing: Georgia Institute of Technology, College of Computing.

Franklin, S. 1998. Artificial Minds. London: MIT press.

Fuchs, C. & Hofkirchner, W. 2005. The Dialectic of Bottom-up and Top-down Emergence in Social Systems. tripleC 1(1), ISSN 1726-670X: 22.

Gibbs, J. P. 1965. Norms: The problem of Definition and Classification. American Journal of Sociology 60: 8.

Gibbs, J. P. 1981. Norms, Deviance and social control: Conceptual matters. New York: Elsevier. ISBN 0444015515

Gilbert, N. 2002. Varieties of Emergence. Paper presented at the Social Agents: Ecology, Exchange, and Evolution Conference Chicago.

Giles, J.; Internet Encyclopaedias go head to head, http://www.nature.com/news/2005/051212/full/438900a.html

Goffman, I. 1983. The Interaction Order: American Sociological Association 1982 Presidential Address. American Sociological Review, 48(1): 1-17.

Goldspink, C. & Kay, R. 2004. Bridging the Micro-Macro Divide: a new basis for social science. Human Relations, 57 (5), ISSN: 0018-7267: 597-618.

Goldspink, C. & Kay, R. 2007. Social Emergence: Distinguishing Reflexive and Non-reflexive modes, AAAI Fall Symposium: Emergent Agents and Socialities: Social and Organizational Aspects of Intelligence. Washington.

Goldspink, C. 2008. Social Self Regulation in On-line Communities: The Case of Wikipedia International Journal of Agent technologies and Systems, 1(1).

Goldspink, C. & Kay, R. 2008a. Agent Cognitive Capabilities and Orders of Emergence: critical thresholds relevant to the simulation of social behaviours', AISB Convention, Communication, Interaction and Social Intelligence. Aberdeen.

Goldspink, C. & Kay, R. 2008b. Agent Cognitive Capability and Orders of Emergence. In G. Trajkovski & S. Collins (Eds.), Agent-Based Societies: Social and Cultural Interactions.

Goldspink, C. & Kay, R. 2009. Autopoiesis and organizations: A biological view of organizational change and methods for its study. In R. Magalhaes & R. Sanchez (Eds.), Autopoiesis in Organizations and Information Systems: Elsevier Science

Habermas, J. 1976. Some Distinctions in Universal Pragmatics: A working paper. Theory and Society, 3(2): 12.

Hayles, K. 1999. How we became post-human: virtual bodies in cybernetics, literature and informatics. Chicago: University of Chicago Press.

Hemetsberger, A. & Pieters, R. 2001. When Consumers Produce on the Internet: An Inquiry into Motivational Sources of Contribution to Joint-Innovation. Paper presented at the the Fourth International Research Seminar on Marketing Communications and Consumer Behavior, La Londe.

Hogg, M. A. & Abrams, D. 1988. Social Identifications: A social psychology of intergroup relations and group processes. London: Routledge. 0-41500694-5

Kallgren, C. A., Reno, R. R., & Cialdini, R. B. 2000. A Focus Theory of Normative Conduct: When Norms do and do not Affect Behavior. PERSONALITY AND SOCIAL PSYCHOLOGY BULLETIN, 26(8): 10.

Kaptelinin, V. & Nardi, B. 1997. Activity Theory: Basic Concepts and Applications, CHI 97.

Kauffman, S. A. 1987. The Evolution of Economic Webs. In P. W. Anderson & et al. (Eds.), The Economy as an Evolving Complex System: Addison-Wesley.

Kauffman, S. A. 1993. The Origins of Order: Self Organization and Selection in Evolution: Oxford University Press.

Kauffman, S. A. & Macready, W. 1995. Technological Evolution and Adaptive Organizations. Complexity, 1(2): 26-43.

Lea, M., Spears, R., & Daphne, d. G. 2001. Knowing Me, Knowing You: Anonymity Effects on Social Identity Processes within Groups. Personality and Social Psychology Bulletin. , 27: 526-537.

Lewis, R. 1997. An ActivityTheory framework to explore distributed communities. Journal of Computer Assisted Learning, 13: 8.

Lih, A. 2004. Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource, 5th International Symposium on Online Journalism. University of Texas, Austin.

Maas, C. J. M. & Hox, J. J. 2005. Sufficient Sample Sizes for Multilevel Modeling. Methodology, 1(3): 86-92.

Misztal, B. 2001. Normality and Trust in Goffman's Theory of Interaction order. Sociological Theory, 19(3): 312-324.

Orton, D. J. & Weick, K. E. 1990. Loosely Coupled Systems: A reconceptualisation. Academy of Management Review, 15(2): 203-223.

Penner, L., Dovidio, J. F., Piliavin, J. A., & Schroder, D. A. 2005a. Pro-social Behavior: Multilevel Perspective. Annual Review of Psychology, 56: 365-392.

Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. 2005b. Prosocial behavior: Multilevel perspectives. Annual Review of Psychology, 56: 365-392.

Pfeil, U., Zaphiris, P., & Ang, C. S. 2006. Cultural Differences in Collaborative Authoring of Wikipedia. Journal of Computer Mediated Communication, 12: 88-113.

Postmes, T., Spears, R., Sakhel, K., & Daphne, d. G. 2001. Social Influence in Computer mediated communication The effects of anonymity on group behavior. Personality and Social Psychology Bulletin, 27: 1243-1254.

Priedhorsky, R., Chen, J., Lam, S. K., Panciera, K., Terveen, L., & Riedl, J. 2007. Creating, Destroying and Restoring Value in Wikipedia, GROUP'07. Sanibel Island, Florida.

Raymond, E. S. 2001. Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary (Revised ed.): O'Reilly Media. 0596001088

Reagle, J.; A Case of Mutual Aid: Wikipedia, Politeness, and Perspective Taking; 18 August 2008, http://reagle.org/joseph/2004/agree/wikip-agree.html

Ricci, A., Omicini, A., & Denti, E. n.d. Activity Theory as a framework for MAS coordination. Bologna: DEIS Universita di Bologna.

Riva, G. & Galimberti, C. 1998. Computer Mediated Communication: Identity and Social Interaction in an Electronic Environment. Genetic, Social and General Psychology Monographs, 124: 434-464.

Rullani, F. 2005. The debate and the community: The reflexive identity concept and the FLOSS community case 32: Santa'Ann School of Advanced Studies.

Sanger, L.; The Early History of Nupedia and Wikipedia: A Memoir, file:///C:/data/surrey/Case%20Studies/wikipedia/early%20history%20of%20nupedia%20and%20wikipedia.htm

Searle, J. R. 1969. Speech Act: An Essay in the Philosophy of Language. Cambridge: Cambridge University Press.

Smith, J. R., Terry, D. J., & Hogg, M. A. 2007. Social Identity and the attitude-behaviour Relationship: Effects of anonymity and accontability. European Journal of Social Psychology, 37, ISSN: 0046-2772: 239-257.

Stiles, W. B. 1992. Describing Talk: A Taxonomy of Verbal Response Modes: Sage. ISBN 0803944659

Surowiecki, J. 2004. The Wisdom of Crowds: Why the Many are Smarter than the Few. London: Abacus.

Terry, D. J. & Hogg, M. A. 1996. Group Norms and the Attitude-Bahaviour Relationship: A Role for group identification. PERSONALITY AND SOCIAL PSYCHOLOGY BULLETIN, 22: 17.

Terry, D. J., Hogg, M. A., & White, K. M. 1999. The theory of planned behaviour: Self-identity, social identity and group norms. British Journal of Social Psychology, 38, ISSN 0144-6665: 225-244.

Thompson, E. 2001. Empathy and Consciousness. Journal of Consciousness Studies, 8(5-7): 1-32.

Wilkinson, D. M. & Huberman, B. A. 2007. Assessing the value of cooperation in wikipedia. Palo Alto: HP Labs.