Research proposal submitted to the FWO by F. Heylighen and F. Van Overwalle:

The social construction of shared concepts: empirical study and computer simulation of a distributed cognitive process

Previous research

Both participating research groups have extensive previous experience in fundamental research, some of it in collaboration with each other, especially in the domain of cognition and social processes. The Lab for Social Psychology has performed a number of experiments and computer simulations, among others on how people place an individual in a particular category, while modifying certain characteristic features of categories for the subjects of the experiment (Van Overwalle & Labiouse, 2003; Van Rooy et al., 2003). In the Centre Leo Apostel theoretical models and computer simulations of knowledge representation, concepts and collective intelligence have been developed and applied over the web (Heylighen, 1999, 2001a,b).

BACKGROUND AND MOTIVATION

Recently, different disciplines have come to regard cognition as a social phenomenon, distributed over a group of individuals. Sociologists have noted long ago that knowledge is a social construction (Berger & Luckman, 1967), however without proposing an explicit model of this process. Management theorists emphasise knowledge management and learning as essentially organisational processes (Senge, 1990). Computer scientists have built models of how collectives, such as ant colonies or swarms of birds, exhibit intelligence that is lacking in the individuals out of which they are constituted (Bonabeau et al., 1998).

These approaches provide a new perspective for the understanding of social and mental organisation, that can be described as "collective intelligence" (Levy, 1997; Heylighen, 1999), or perhaps more accurately as "distributed cognition" (Hutchins, 1995). While most social sciences focus on the "political" dynamics of competition, cooperation and forming of coalitions, the distributed cognition approach stresses the social system as a creator and carrier of practical knowledge (Gaines, 1994). There are immense potential applications of this approach for the design of organisational entities that can cope with situations which are too complex for the limited cognitive capacity of a single individual. Yet, the distributed cognition approach lacks a unified foundation, and is as yet little more than a heterogeneous collection of ideas, observations and techniques, coming from a variety of disciplines and traditions.

AIM

The present research proposal aims to address the issues of distributed cognition at the most fundamental, transdisciplinary level, thus laying the foundation for a sound theoretical approach underlying the empirical work. The building blocks of knowledge are concepts, i.e. the categories into which our mind classifies perceptions (Lakoff, 1987; Lamberts & Shanks, 1997). By associating and combining concepts we describe, understand and design reality. The social construction of knowledge, therefore, requires in the first place the social development of concepts, i.e. a group of individuals reaching a consensus on how to categorise observed phenomena. Even though primitive perceptual categories such as colours (e.g. "red", "green", "blue") are determined partly biologically, partly individually, socio-cultural processes play a substantial part in their creation, as illustrated by the different colour categorisations used in different languages (Belpaeme, 2001). Social construction has an even greater role in the development of more abstract concepts, such as "democracy" or "intelligence".

In this project we intend to investigate the fundamental socio-cognitive processes through which shared concepts are constructed out of individual concepts. We shall focus in particular on the factors that influence this process and on the manner in which the resulting "consensual" concept differs from the initial individual concepts. That, we hope, will give us a better insight into the mechanisms guiding distributed knowledge development, which in turn will allow us to increase the efficiency and reliability of that process.

A crucial application of our research is the design of formal "ontologies" (Staab & Studer, 2003), i.e. the systems of categories necessary for knowledge representation in the "semantic web" (Berners-Lee et al., 2001). This knowledge architecture for the future Internet allow us to get unambiguous answers to specific questions, in sharp contrast to the long lists of often rather less than more relevant documents that we usually get from present search engines (Heylighen, 2001b). Ontologies are compiled by committees of experts, but as yet no methodology exists that helps such committees to reach consensus. The danger is that the ontologies constructed by the most powerful players (e.g. Microsoft) will become dominant, rather than those beneficial to society as a whole.

RESEARCH HYPOTHESES

We start with a number of initial hypotheses, inspired by previous research (Heylighen, 1999; Van Overwalle et al., 2003). While there is a divergence of theories about the internal, mental structure of a concept (Lamberts & Shanks, 1997), our "distributed" perspective focuses only on its "external" use. We operationalise a concept as a process of categorisation, whereby different phenomena are classified as instances of this concept to a greater or lesser degree. The colour of blood, for example, will be classified with certainty (strength 1) as "red"; that of a brick - with a strength 0.7; that of an orange - with a strength 0.3; that of grass with 0. A concept can thus be represented as a vector, e.g. (1, 0.7, 0.3, 0), the components of which correspond to the categorisation strengths. Such representations in multidimensional vector spaces have proven their usefulness in the semantic analysis of concepts (see Heylighen, 2001b, Foltz, 1996).

Concepts are created as abstractions of recurrent aspects of reality. Each individual or agent will experience the world from its own perspective, encountering different phenomena at different moments. Therefore, the abstractions made by different agents will in general be different as well, in the sense that the categories will not completely overlap. For example, a child that has just learned the concept "cat" by abstracting invariant features such as "small", "furry", "four legs" from a few instances, when it first encounters a dog will in general view this creature as a somewhat unusual instance of the "cat" category, until its parents (social consensus) point out the discrepancy.

In order to communicate effectively, different agents must use the same categories. Otherwise misunderstandings will arise, e.g. when one individual asks another one to paint something in "red", but is unhappy with the result because it appears to him as "brick". To achieve effective coordination, agents must reach a shared understanding of a concept, so that they agree about which phenomena are or are not instances of the concept. This requires a process of negotiation or cognitive interaction.

Our hypothesis is that a group of individuals interacting in this way will undergo a process of self-organisation (Steels, 1998; Bonabeau et al., 1998), whereby out of local interactions a global, more coherent pattern emerges. This implies that the divergence in categorisations among the agents will diminish, ideally leading to a single concept. This shared concept will also be qualitatively "better" (more general, more reliable, more easy to use...) than the initial concepts, as it will integrate the diversity of subjective experiences in a broader, intersubjective whole. Johnson's (1998) computer simulation has already demonstrated that even a simple "averaging" of individually acquired knowledge by agents in a group produces greater reliability, even though it did not take into account the non-linear effects of self-organisation (Heylighen, 1999).

We intend to operationalise this hypothesis in the following way. For each subject, the individual concept is represented by a vector. The comparison of the different vectors gives us an objective measure for the spread or diversity in the viewpoints. The average of the vectors defines the "collective" concept for the group (Heylighen, 1999). After the subjects have interacted, individual and averaged concepts are measured again. Based on our hypothesis we expect the following to hold true: 1) the spread among the subjects will diminish; 2) the collective concept will be consolidated, in the sense that vector components about which there was a relative agreement will be strengthened, while components important to only one or very few individuals are reduced, or disappear altogether; 3) in the cases where the opinions diverge strongly we foresee an alternative scenario in which the non-linear dynamics leads to polarisation splitting the vectors into two or more clusters that could be viewed as alternative interpretations of the concept.

These hypotheses will be tested and developed into a more detailed model by investigating the factors that control the process. At least the following factors are likely to be relevant: diversity among the subjects, unicity of their perspectives, type of interaction, generality or ambiguity of the concept. A better understanding of these elements and their causal effect will allow us to choose them in such a way as to maximise the quality of the consensual concept.

DESIGN AND METHODOLOGY

Work on collective intelligence has typically relied on computer simulations, in which interacting software agents form an "artificial society" (Bonabeau et al., 1998). This approach has the advantage that many variations of a basic model can be tested quickly and easily, e.g. by varying the parameters of the simulation. The disadvantage is that a simulation is a very simplified model of reality, which is wholly dependent on the subjective assumptions of the designer. Real-life observations of actual social systems, as used in the distributed cognition tradition (Hutchins, 1995), can evade these criticisms, by providing an open-ended source of unanticipated effects and interactions. The disadvantage is that they are very time-consuming and difficult to control so that only a few variations of a basic situation can be investigated. The present project therefore wishes to combine the benefits of both methodologies, using observation to suggest new hypotheses and simulation to quickly explore the different implications of these hypotheses, so that the most promising ones can become the focus of a new observation.

EMPIRICAL APPROACH

In our basic set-up, a small group (about 10) of experimental subjects are requested to discuss a given concept, with the objective of achieving a shared understanding. The concept is chosen such that everyone has some experience with it, but there remains sufficient vagueness or ambiguity to allow different interpretations. To minimize the risk for emotional arguments or political games, the concepts are selected to be as neutral as possible (e.g. "system", "idea", "fruit"), and the participants are told explicitly that there won't be any "winners" or "losers". The subjects are informed about the subject before the experiment, so that they can prepare their thoughts without mutually influencing each other. They are asked in particular to suggest a number of examples, counterexamples and intermediate cases of the category. We select the most representative ones of those, and submit the resulting list of some thirty items to all subjects. We ask them to score each one on a 10-point scale, indicating the degree to which they consider it to belong to the category. This produces the initial concept vectors for all participants.

In the written version of the experiment, the interaction takes place asynchronously, using the AWE electronic discussion system developed at the Centre "Leo Apostel" (http://pcp.vub.ac.be/riegler/awe/). Each participant starts with a short description of what the concept means for him or her, and then is allowed to reply to the interpretations of others, using examples, arguments and counterarguments. After a period long enough to allow each subject to intervene several times, the discussion is stopped, and the concept vectors are measured again. The statistical comparison of initial and final vectors provides us with a quantitative analysis of the evolution of the concept. A textual analysis of the different interventions provides us with a more qualitative picture of the arguments and factors that have influenced the outcome. The possible reasons why a particular participant has or has not changed positions are explored by focused interviews.

The oral version of the experiment is similar, except that the group of participants now discuss face-to-face during a two hour session without facilitation. Concept

vectors are again measured before and after the session. The discussion is recorded on videotape, and afterwards analysed for specific factors that appear to have influenced the outcome. Immediately after the session, selected participants are interviewed in order to explore their unstated reasons for changing their perspective.

SIMULATION

The simulation model will start from the KEBA virtual environment in which autonomous software agents learn concepts by abstracting invariant patterns from the sequence of their perceptions and actions (Gershenson, 2002). This environment is extended with distributed cognitive interactions, inspired by recent simulations of the social construction of symbols or concepts (e.g. Hutchins & Hazelhurst, 1995). Agents interact according to the following protocol (similar to what Steels (1998) and others call a "language game"):

Two agents "meet" in virtual space, which means that they perceive the same virtual phenomena (e.g. a rock, a piece of food) from the same perspective. The first agent indicates a phenomenon that belongs to a category learned earlier, and the second one finds the best fitting category for that phenomenon in its cognitive system. The second agent now in turn indicates a phenomenon belonging to that same category. If this phenomenon also belongs to the same category for the first agent, both categorisations are reinforced, otherwise they are weakened. In the next move of the "game", another phenomenon is indicated which may or may not belong to the category, and the corresponding vector component is strengthened or weakened depending on the degree of agreement. After a number of moves the game is stopped, each agent maintaining the mutually adjusted categories. Each agent in turn is randomly coupled to another agent in the collective, in order to play a new game using different phenomena.

According to our hypotheses, after a sufficient number of games a stable and coherent system of categories shared by all agents will emerge through self-organization. (A similar set-up can be found in Belpaeme's (2001) simulation of the origin of shared colour categories, the difference being that our approach does not take into account the "words" that an agent choose to express a concepts, but does take into account the individual experience in learning concepts from the environment). The simulation will be iterated for a large number of values of the relevant factors (e.g. number of movers per game, number of games, number of agents, diversity of agents, homogeneity of categories, etc.), in order to determine their influence as accurately as possible.

COORDINATION BETWEEN THE PARTICIPANTS

The theoretical model and the computer simulation will be developed at the Centre "Leo Apostel" under the supervision of F. Heylighen. F. Van Overwalle, at the Social Psychology Laboratory, will supervise the experiments, and help with the development of the model and simulation. Both approaches will continuously interact, the one providing feedback and suggestions for deeper exploration to the other.

In a first stage (2004), the theoretical model is elaborated on the basis of the literature and discussion with experts, the simulation is prepared, and about six written and oral experiments are held as pilot studies. The most relevant factors are distilled out of the results. In the second stage (2005-2006), a number of experiments and

simulations are executed in parallel, controlling each of these factors individually (e.g. small vs. large groups, general vs. specific concepts). In the last stage (2007) all results are interpreted and synthesised, and published in the form of PhD theses and peer-refereed papers.

REFERENCES

Belpaeme T. (2001) Reaching coherent colour categories through communication. In Kröse, B. et al. (eds.), Proc. 13th Belgium-Netherlands Conference on AI, Amsterdam, p. 41-48.

Berger P. L., T. Luckmann: (1967) The Social Construction of Reality: A Treatise in the Sociology of Knowledge, Anchor.

Berners-Lee ,T., J. Hendler & O. Lassila (2001): The Semantic Web, Scientific American, 282(5),

Bonabeau, E., Dorigo, M. & Theraulaz, G. (1998). Swarm Intelligence, Oxford University Press

Foltz P.W. (1996) Latent Semantic Analysis for text-based research, *Behavior Research Methods, Instruments, & Computers*, 28, 197-202.

Gaines B.R. (1994), The Collective Stance in Modeling Expertise in Individuals and Organizations, *Int. J. Expert Systems* 71, 22-51.

Gershenson, C., (2002), Behaviour-based Knowledge Systemse, in: Proc. 2nd Int. Workshop on Epigenetic Robotics. Edinburgh.

Heylighen F. (1999): Collective Intelligence and its Implementation on the Web, *Computational and Mathematical Theory of Organizations* 5, p. 253-280.

Heylighen F. (2001a): Bootstrapping knowledge representations, *Kybernetes* 30, p. 691-722.

Heylighen F. (2001b): Mining Associative Meanings from the Web:, in: *Proc. Int. Colloquium: Trends in Special Language & Language Technology*, R. Temmerman & M. Lutjeharms (eds.) (Standaard, Antwerpen), p. 15-44.

Hutchins E (1995): Cognition in the Wild (MIT Press).

Hutchins, E. & B. Hazelhurst (1995). How to invent a lexicon: the development of shared symbols in interaction. In N. Gilbert and R. Conte (Eds.), Artificial Societies. UCL Press

Johnson N. (1998) Collective Problem-Solving (Los Alamos National Laboratory technical report).

Lakoff G. (1987) Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. University of Chicago Press.

Lamberts K. & D. Shanks (eds.) (1997): Knowledge, concepts and categories, MIT Press.

Lévy P. (1997): Collective Intelligence, Plenum.

Senge P. (1990) The Fifth Discipline: the Art and Practice of Learning Organizations, Doubleday.

Staab S. & Studer R., (eds.) (2003), Handbook on Ontologies in Information Systems, Springer Verlag.

Steels L. (1998): Synthesising the origins of language and meaning using co-evolution, self-organisation and level formation, in Hurford et al. (eds): Approaches to the evolution of language (Cambridge University Press), p. 384-404.

Van Overwalle, F., & Labiouse, C. (2003) A recurrent connectionist model of person impression formation. Personality and Social Psychology Review, in press.

Van Rooy, D., Van Overwalle, F., Vanhoomissen, T., Labiouse, C. & French, R. (2003). A recurrent connectionist model of group biases. Psychological Review, in press.