submitted for Proceedings of Collective Intentionality IV, Siena (Italy), to be published as a special issue of Cognitive Systems Research

The Emergence of Distributed Cognition: a conceptual framework

Francis HEYLIGHEN, Margeret HEATH and Frank VAN OVERWALLE

Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium
{fheyligh, mheath, fjvoverw}@vub.ac.be

ABSTRACT: We propose a first step in the development of an integrated theory of the emergence of distributed cognition/extended mind. Distributed cognition is seen as the confluence of collective intelligence and "situatedness", or the extension of cognitive processes into the physical environment. The framework is based on five fundamental assumptions: 1) groups of agents self-organize to form a differentiated, coordinated system, adapted to its environment, 2) the system co-opts external media for internal propagation of information, 3) the resulting distributed cognitive system can be modelled as a learning, connectionist network, 4) information in the network is transmitted selectively, 5) novel knowledge emerges through non-linear, recurrent interactions. The implication for collective intentionality is that such a self-organizing agent collective can develop "mental content" that is not reducible to individual cognitions.

Extended Mind: collective intelligence and distributed cognition

From a cybernetic perspective [Heylighen & Joslyn, 2001], a cognitive system cannot be understood as a discrete collection of beliefs, procedures, and/or modules. Cognition is a continuously evolving process which relates present perceptions via internal states to potential further perceptions and actions. It thus allows an agent to anticipate what may happen, adapt to changes in its environment, and moreover effect changes upon its environment [Kirsch & Maglio, 1994].

The study of cognition—cognitive science—is in essence multidisciplinary, integrating insights from approaches such as psychology, philosophy, artifical intelligence (AI), linguistics, anthropology, and neurophysiology. To this list of sciences of the *mind*, we now also must add the disciplines that study *society*. Indeed, an increasing number of approaches are proposing that cognition is not limited to the mind of an individual agent, but involves interactions with other minds.

Sociologists have long noted that most of our knowledge is the result of a social construction rather than of individual observation [e.g. Berger & Luckman, 1967]. Philosophers have brought the matter to research for urgent consideration in theories of mind [e.g. Searle, 1995]. The nascent science of memetics [Aunger, 2001; Heylighen, 1998], inspired by evolutionary theory and culture studies, investigates the spread of knowledge from the point of view of the idea or *meme* being communicated between individuals rather than the individual that is doing the communication. Economists too have started to study the role of knowledge in innovation, diffusion of new products and technologies, the organization of the market, and overall social and economic development [Martens, 2004]. Management theorists emphasise knowledge management and learning as an organisational phenomenon rather than as an individual process. Effective organisational learning is deemed to be the difference between an enterprise that flourishes and one that fails [Senge, 1990]. Social psychologists have started to do laboratory experiments to study cognition at the group level [e.g. Brauer et al., 2001; Klein et al., 2003; Van Rooy, Van Overwalle et al., 2004]. Biologists, aided by computer scientists, have built models that demonstrate how collectives of simple agents, such as ant colonies, bee hives, or flocks of birds, can process complex information more effectively than single agents facing the same tasks [Bonabeau et al., 1999]. Building on the tradition of distributed artificial intelligence, the subject of collective cognition is now even being investigated mathematically [Crutchfield et al. 2002].

These different approaches provide a new focus for the understanding of cognition that might be summarized as *collective intelligence* [Levy, 1997; Heylighen, 1999], i.e. the cognitive processes and structures that emerge at the social level. But at the same time the investigation of cognition has expanded in another direction: that of the physical environment.

The failure of traditional, “symbol-processing” AI to come up with workable models of intelligence has pointed to the necessity for *situatedness, embodiment* or *enaction* [Steels & Brooks, 1995; Clark, 1997]. This refers to the observation that cognition or mind cannot exist in a mere abstract realm of ideas (the “brain-in-a-vat”), but must be part of an interaction loop, via perception and action, with a concrete environment [cf. Heylighen & Joslyn, 2001]. This has led to a flurry of interest in autonomous robots which forego complex representations and symbol manipulations by using the environment as its own best model [Steels & Brooks, 1995].

The environment supports cognition not just *passively*—by merely representing itself, but *actively*—by registering and storing agent activities for future use, and thus functioning like an external memory [Kirsh, 1996; Kirsh & Maglio, 1994; Clark, 1997]. Examples abound, from the laying of pheromone trails by ants and the use of branches to mark foraging places by wood mice to the notebooks we use to record our thoughts. Physical objects can further be used to collect and process information, as illustrated by telescopes and computers.

This “offloading” of information onto the environment makes this information potentially available for other agents, thus providing a medium by which information sharing, communication, and coordination can occur. This basic mechanism, known as “stigmergy”, underlies many examples of collective intelligence [Clark, 1997; Heylighen, 1999; Susi & Ziemke, 2001], such as the trail laying of ants and the mound building of termites. Thus, the idea of an *active externalism* [Clark & Chalmers, 1998] and the creation of *epistemic structure* in the environment by the cognizer [Kirsch & Maglio, 1994] may provide a foundation for the perspective of *distributed cognition* [Hutchins, 1995], demonstrating intentional computational interplay between human agents, technology and environment. This makes for a strong case for collective intentionality under the umbrella of the extended mind thesis.

Extending the extended mind: towards an integrated theory

The question still remains: how is it possible that a world, fundamentally characterized by ‘flex or slop’ in its material nature [Cantwell Smith, 1996] can ever be brought into intentional coordination by something like an extended mind? Moreover, can the extended mind thesis as it stands, rather than as we see it being extended, explain collective intentionality, where intentionality is defined as “a way of exploiting local freedom or slop in order to establish coordination with what is beyond effective reach” [Cantwell Smith, 1996, p. 208]?

Let us extend the functionalism implied in the extended mind thesis in a cybernetic sense, so that it includes a theory of how this organisation comes about, how these functional relationships arise in the first sense rather than just how they survive, what would constitute a unit of analysis for distributed cognition and how a theory of computational intentionality could explain our proposal. The extended mind thesis could then be seen as the interlinking of multiple brains, forming kinds of associative engines where their environmental interactions are iterations of series of simple pattern completing- or pattern-creating real world actions (computations). Furthermore we would propose that this extension of the extended mind thesis would complement Cantwell Smith's assertions [1996, p. 207] that “coordinating regularity (what will ultimately become the semantic relation) and the coordinated-with regularity (what will ultimately become the type structure of the referent) will emerge together”. Cantwell Smith's assertion that it is this very ‘*flex and slop*’ in the material world which gives rise to the requirement for coordination, is best explained by a theory of computational intentionality, and to our minds a theory best extended by extending the extended mind thesis.

Hutchins [1995] exemplifies this approach in part by his ethnographic studies of distributed cognition as "the propagation of representational states across representational media". This study identifies computational distinctions of the system as a whole versus those of an individual task activity. However, in spite of its promises distributed cognition as yet offer little more than a heterogeneous collection of ideas, observation techniques, preliminary simulations and ethnographic case studies. It lacks a coherent theoretical framework that would integrate the various concepts and observations, and provide a solid foundation that would allow researchers to investigate varieties of cognitive systems.

For us, getting back to basics means understanding how distributed cognition emerges and is generated. The analysis of an existing cognitive process, such as ship navigation [Hutchins, 1995], is not sufficient, because the underlying distributed systems tend to be complex and specialized, while their often convoluted way of functioning is typically rigidly set as the result of a series of historical accidents. A more general and fundamental understanding, not only of the "how?" but also the "what?" and the "why?", may be found by analysing how distributed cognition emerges step by step from a system that initially does not have any cognitive powers. We wish to focus on the creation—and not merely the propagation—of information in these systems.

In examining a collective intentionality, our basic research questions could be formulated as follows: How do initially independent agents, interacting by means of external media come to form an intentional cognitive system? What kind of coordination between different computational structures creates a distributed cognitive system and one which exemplifies properties associated with mind? Which features influence the efficiency of the process by which this happens? For example, do resulting cognitive capabilities, inherent in these systems depend on the number of agents, the sequencing of information activation, the diversity of agents, the presence or absence of different types of media?

Potential applications

An extended theory of the extended mind as we we envisage it here (in form of a mechanism of distributed cognition) would offer a wealth of potential applications. To start with, understanding how knowledge and information are distributed throughout social systems would help us to foster the economic and social development that new knowledge and better coordination engenders [Martens, 1998; 2004]. In particular, such a theory should tell us how important new ideas can diffuse most efficiently, and conversely how the spread of false rumours, superstitions and "information parasites" might be curtailed [Heylighen, 1999]. More generally, it may help us to control for the cognitive biases and social prejudices whose ubiquity psychologists have amply demonstrated [Brauer et al., 2001; Klein et al., 2003; Van Rooy, Van Overwalle et al., 2004].

On a smaller scale, a theory of distributed cognition has immediate applications in business, government, and other organizations. It would help them to promote innovation and avoid the pitfalls of collective decision-making, such as *groupthink* [Janis, 1972], which stifle creativity. It would support organizations not only in generating new knowledge but in efficiently maintaining, applying and managing the knowledge that is already there. More fundamentally, it would provides us with concrete guidelines to design more effective organizations, where roles and functions are clearly specified, and where information is processed in a coordinated way, with a minimum of loss, distortion, misunderstanding or confusion. In sum, it would foster the collective intelligence of the organization, while minimizing the inherent tendency of groups towards "collective stupidity".

Technological applications abound as well. A crucial application of the proposed model of distributed cognition would be the compilation by committees of experts of formal "ontologies" [Staab & Studer, 2003], i.e. the systems of categories necessary for the *semantic web* [Berners-Lee et al., 2001]. This knowledge architecture for the future Internet will allow users to get concrete answers to specific questions, while enabling various services to coordinate automatically. But this requires efficient and consensual schemes to represent knowledge that is generated and managed in a distributed manner. More generally, a lot of research is going on in distributed AI to develop efficient coordination schemes to let software agents collaborate. One of the more immediate application domains is *ambient intelligenc*e [ISTAG, 2003]. This refers to the vision of everyday artefacts and devices such as mobile phones, coffee machines and fridges exchanging information and coordinating with each other so as to provide the best possible service to the user, without needing any programming or prompting—thus effectively extending the user's mind into his or her physical environment [Gershenson & Heylighen, 2004].

Integrating the ambient intelligence of devices, the collective intelligence of organizations and society, and the global communication and coordination medium that is the future Internet leads us to a vision of a *global brain* [Heylighen, 1999; Heylighen & Heath, 2004], i.e. an intelligent network formed by the people of this planet together with the knowledge and communication technologies that connect them together. This vision of a global brain is in fact the ultimate extension to an extended mind theory. It is to this end that we direct our research in intentionality.

Assumptions for building an integrated theory

Inspired by our earlier research, we wish to propose five fundamental "working hypotheses", which can function as starting points or postulates for building a general model of (i) distributed cognition, (ii) collective intelligence and (iii) extended mind.

1. groups of agents self-organize

Consider a group of initially autonomous actors, actants or agents, where the agents can be human, animal, social or artificial. Agents by definition perform *actions*. Through their shared environment the action of the one will in general affect the other. Therefore, agents in proximity are likely to *interact*, meaning that the changes of state of the one causally affect the changes of state of the other. These causal dependencies imply that the agents collectively form a *dynamical system*, evolving under the impulse of individual actions, their indirect effects as they are propagated to other agents, and changes in the environment. It is important to note that a dynamical system has computational structure and is therefore able to process information. Not only that, but the dynamics themselves will generate a pattern, not just seek to complete it [Crutchfield, 1998]. Moreover, this system will typically be non-linear, since causal influences normally propagate in cycles, forming a complex of positive and negative feedback loops.

While such a complex system is inherently very difficult to model, control or predict, all dynamical systems tend to self-organize [Heylighen & Joslyn, 2001; Heylighen, 2003; Heylighen & Gershenson, 2003], i.e. evolve to a relatively stable configuration of states (an *attractor* of the dynamics). In this configuration, we can say that the agents have mutually adapted [Ashby, 1962], restricting their interactions to those that allow this collective configuration to survive. There is moreover an on-going selective pressure to make these interactions more synergetic [Wright, 2000], because a mutually beneficial interaction is preferable to one that is less so. In this view, the self-organization and further evolution of the collective effectively create a form of "social" *organization*, in which agents help each other so as to maximize the collective benefit, as illustrated by the many simulations of the evolution of cooperation [e.g. Axelrod, 1984; Riolo, Cohen & Axelrod, 2001; Hales & Edmonds, 2003].

An effective, synergetic organization requires a coordinative tissue of action. According to the classification scheme of *coordination theory* [Crowston, 2003], we can distinguish the following fundamental dependencies between activities or processes: 1) two processes can use the same resource (input) and/or contribute to the same task or goal (output); 2) one process can be prequisite for the next process (output of the first is input of the second). If we associate the activities to agents, the first case calls for tasks to be performed in parallel and the second case in sequence. Efficient organization means that the right activities are delegated to the right agents at the right time. The parallel distribution of tasks determines the *division of labor* between agents. The sequential distribution determines their *workflow*.

Division of labor reinforces the specialization of agents, allowing each of them to develop an expertise that the others do not have [Gaines, 1994; Martens, 2004]. This enables the collective to overcome individual cognitive limitations, accumulating a much larger amount of knowledge than any single agent might. Workflow allows information to be propagated and processed sequentially, so that it can be refined at each stage of the process. Self-organization thus potentially produces emergent cognitive capabilities that do not exist at the individual level. Moreover, it may give rise to unexpected organisational properties such as the emergence of a requirement of a new function, the loss of crucial information, the development of additional tasks and the deviation from existing workflow rules [Hutchins 1995].

2. the system co-opts external media for communication

Self-organization in this sense can be seen as the more efficient, synergetic use of interactions. Interactions between agents necessarily pass through their shared physical environment. We will call the external phenomena that support these interactions *media*. Certain parts or aspects of the environment better lend themselves to synergetic interaction than others do. For example, a low-bandwidth communication channel that is difficult to control and that produces a lot of errors, such as smoke signals, will support less synergetic interactions than a reliable, high-bandwidth one, such as optical cable. Thus, there is a selective pressure for agents to preferentially learn to use the more efficient media, i.e. the ones through which causal influences—and therefore information—are transmitted most reliably and accurately.

Moreover, simply by using them, the agents will change the media, generally adapting them to better suit their purposes. For example, animals or people that regularly travel over an irregular terrain between different target locations (such as food reserves, water holes or dwellings) will by that activity erode paths or trails in the terrain that facilitate further movement. The paths created by certain agents will attract and steer the actions of other agents, thus providing a shared coordination mechanism that lets the agents communicate indirectly (stigmergy). A slightly more sophisticated version of this mechanism are the trails of pheromones laid by ants to guide other members of their colony to the various food sources, thus providing them with a *collective mental map* of their surroundings [Heylighen, 1999]. Humans, as specialized tool builders, excel in this adaptation of the environment to their needs, and especially in the use of physical supports such as paper, electromagnetic waves or electronic hardware to store, transmit and process information.

The evolutionary origin of such externally mediated communication can be understood by noting that agent actions (e.g. moving, eating, drinking, ...) will in general leave some kind of side-effects or *traces* in their shared environment. Some of the traces may remain for a long time (e.g. paths eroded in rocky terrain), others will be very short-lived (e.g. a cry of anguish). We assume that agents can perceive basic phenomena in their environment, including other agents' traces, and that they learn to associate these features with other features and with their in-built goals (e.g. finding food). They thus will learn to recognize which traces provide useful information about the phenomena that are important to them (e.g. food).

From this individual adaptation, there seem to be two alternative paths for inter-individual evolution:

1) The trace is beneficial for the agent that perceives it (e.g. pointing a predator towards its prey), but detrimental to the one that made it (e.g. making the prey more visible for the predator). In that case we can expect an arms-race type of evolution, in which "predators" become better at detecting traces, while "prey" agents become better at hiding their traces. This is unlikely to lead to any kind of shared medium.

2) The trace is useful for both parties (for example because it indicates a shared danger). In this case, there will be a selective pressure for both parties to make the trace easier to perceive, by becoming more adept a leaving clear, stable and informative traces and at distinguishing and interpreting traces left by others. Thus, the trace will co-evolve with the agents' cognitive abilities, to become an efficient, shared communication medium that allows one agent to leave messages for itself and others.

In this way, external media are increasingly assimilated or co-opted into the social organization, making the organization's functioning ever more dependent on them. As a result, the cognitive system is extended into the physical environment and can no longer be separated from it.

3. distributed cognitive systems function like connectionist networks

We can represent an extended social organization as an abstract network as follows: we assign to nodes the function of agents or objects that store or contain information, and to the links that connect them the channels along which information is communicated (we could also argue for media as nodes and agents as links). Links can have variable strength, where strength represents the ease, frequency or intensity with which information is transmitted. They represent stabilized causal influences between agents and/or objects, possibly supported by co-opted media.

Every node is characterized by its space of possible states. The actual state at the beginning of a process is propagated in parallel along the different links across the different media, and recombined in the receiving nodes. State spaces can in general be factorized into independent variables or degrees of freedom, each of which can take on a continuum of values [Heylighen, 2003]. A complex node can thus be functionally decomposed as an array of simple, one-dimensional nodes that only take on a single "intensity" or "activation" value. The resulting network of simple nodes and links seems functionally equivalent to a "neural" or *connectionist* network, where activation spreads from node to node via variable strength links [Van Overwalle & Labiouse, 2004; McLeod et al., 1998]. This network is in general recurrent, because of the existence of cycles or loops as mentioned earlier.

Connectionist networks have proven to provide very flexible and powerful models of cognitive systems [Van Overwalle & Labiouse, 2004; McLeod et al., 1998]. Their processing is intrinsically parallel and distributed. Because of the accompanying redundancy, they are much more robust than purely sequential architectures, surviving destruction of part of their nodes and links with merely a "graceful" degradation of their performance. As a result, these systems do not need a central executive, eliminating the need for centralized and deliberative processing of information. Moreover, since activation spreads automatically to other nodes than those that received the initial stimuli, connectionist networks exhibit emergent properties such as pattern completion and generalization, allowing lacking data to be filled in, and inferring plausible conclusions on the basis of very limited information.

Most importantly, connectionist networks inherently support learning, by means of the continuous adaptation of the link strengths to the ways in which they are used. Thus, succesfully used links become stronger, making it easier for information to be propagated along them, while links that are rarely used or whose use led to erroneous results, weaken. In an extended cognitive system we can conceive of at least two mechanisms for such a reinforcement or inhibition. In the material sense, as proposed in the previous hypothesis, commonly used media become more effective. But a more flexible mechanism is social adaptation, in which an agent learns from the experience of communicating with another agent. If the other agent reacts appropriately, the first agent will increase its trust in the other's competence and goodwill, and thus becomes more likely to communicate similar information to that agent in the future.

As such, the network's "experience" of use is stored in long-term weight changes of the connections. Thus, the network acquires new knowledge in a distributed manner, i.e. storing it in the pattern of links rather than in the states or memories of individual nodes.

4. information in the network is propagated selectively

Whether information is transmitted will not only depend on the architecture of the network, but on the content of the information. Memetic analysis and social-psychology observation have suggested different selection criteria that specify which information is preferentially passed on [Heylighen, 1997, 1998]. These include the criteria of:

- *utility* (the information is useful or valuable to the agents)
- *novelty* (the information is not already known)
- *coherence* (the information is consistent with the knowledge that the agents already have)
- *simplicity* (since complex information is difficult to process, less important details tend to be left out)
- *formality* (the less context or background communicating agents share, the more important it is to express the information explicitly)
- *expressivity* (the information is easily expressible in the available media)
- *authority* (the source is recognized as being trustworthy)
- *conformity* or *consensus* (the majority of agents agree on the information)

Several of these criteria have been empirically confirmed through psychological experiments [Lyons & Kashima, 2003] and analysis of linguistic data [Heylighen & Dewaele, 2002; Chielens, 2003]. They provide a simple set of guidelines to understand the evolution of distributed knowledge through variation and selection [Heylighen, 1998].

A theory of distributed cognition would ideally allow these criteria to be derived from the dynamics of a distributed connectionist network, rather than have them posited to some degree *ad hoc*. A preliminary simulation [Van Overwalle, Heylighen & Heath, 2004] indeed suggests that this can be achieved. For example, the reinforcement of links through the increase of trust builds authority for the sending agents, while telling them which information the receiving agents are likely to already know and agree with, making it less important for them to transmit detailed, explicit reports. Moreover, spread of activation along existing connections will automatically attenuate inconsistent or complex signals, while amplifying signals that are confirmed by many different sources (conformity) or that activate in-built rewards or punishments (utility).

Selective propagation and thus filtering out of less relevant or less reliable data already constitutes information processing, as it compresses the data and thus potentially distils the underlying pattern or essence. However, if selectivity is inadequate, this can lead to the loss of important ideas, and the propagation of incorrect information, as exemplified by the flurry of social and cognitive biases that characterizes "groupthink" [Van Rooy, Van Overwalle, Vanhoomissen et al., 2003]. More extensive modelling and simulation should allow us to identify the central factors through which we can control these dangerous tendencies.

5. novel knowledge emerges

In extending the extended mind hypothesis in the way that we have, our notion of multiple brains as associative engines is well served by connectionist models of social systems. Groups often can be more intelligent than individuals, integrating information from a variety of sources, and overcoming the individual biases, errors and limitations. In the simplest case, this occurs through a superposition of individual contributions. Because of the law of large numbers, the larger the variety of inputs, the smaller the overall effect of random errors, noise, or lacking data, and the clearer and more complete the resulting collective signal [Heylighen, 1999]. This "averaging" of contributions is represented very simply in a connectionist network, by the activation from different inputs being added together and renormalized in the target nodes.

But a recurrent connectionist network, being non-linear and self-organizing, may offer more radical forms of novelty creation, through the emergence of structures that are more than the sum of their parts. Rather than being attenuated by averaging, noise can here play a creative role, triggering switches to a wholly new attractor or configuration at the bifurcation points of the dynamics, thus exemplifying the "order from noise" principle [Heylighen, 2003].

The same mechanisms of self-organization that may lead to coordination between agents are also likely to lead to coordination and integration of the ideas being communicated between those agents. An idea that is recurrently communicated will undergo a shift in meaning each time it is assimilated by a new agent, who adds its own, unique interpretation and experience to it. Moreover, the need to express it in a specific medium will also affect the shape and content of the idea, which will be further constrained by the need to achieve a shared reference of intentionality for it. Like in a game of Chinese whispers, by the time the idea comes back to the agent who initiated it, it may have changed beyond recognition. After several rounds of such passing back and forth between a diverse group of agents, the dynamical system formed by these propagations with a twist is likely to have reached an attractor, i.e. an invariant, emergent configuration.

In this way, novel shared concepts may self-organize through communication, providing a basic mechanism for the social construction of knowledge [Berger et al., 1967]. Concrete illustrations of this process can be found in multi-agent simulations of the origin of language where the symbol (external support) co-evolves with the category that it refers to (internal concept with external reference) [e.g. Hutchins & Hazelhurst, 1995; Steels, 1998; Belpaeme, 2001]. These models are based on recursive *language games*, where a move consists of one agents expressing a concept and the receiving agent indicating whether or not it has "understood" what the expression refers to (e.g. by pointing towards a presumed instance of the category), after which the first agent adjusts its category and/or expression. After a sufficient number of interaction rounds between all the agents in the collective, a "consensus" typically emerges about a shared concept and its expression. Thus, such models may provide a first account of the emergence of collective intentionality as a distributed, self-organizing process.

Knowledge consists not only of concepts or categories, but of associative, logical or causal *connections* between these categories. These have the general form:

IF occurrence of category *A* (e.g. *banana* or *lack of preparation*),
THEN expect occurrence of category *B* (e.g. *yellow* or *failure for exam*).

Such basic connections underlie not only expectation or prediction, but causal attribution or explanation of *B*, given *A*. The connections between categories can be learned through the closely related *Hebbian* [e.g. Heylighen & Bollen, 2002] or *Delta algorithms* [Van Overwalle & Labiouse, 2004]. These connectionist learning rules are simple and general enough to be applicable even when cognition is distributed over different agents and media [e.g. Heylighen & Bollen, 2002; Van Overwalle, Heylighen & Heath, 2004]. However, if we moreover take into account the social construction of concepts, we get a view of concepts, symbols, media and the connections between them co-evolving, in a complex, non-linear dynamics. This points us towards a potential "bootstrapping" model of how complex and novel distributed cognitive structures, such as languages, scientific theories, world views and institutions, can emerge and evolve.

Implications for collective intentionality

Intentionality in its philosophical sense denotes the assumption that all mental phenomena, such as experiences, beliefs, conceptions or desires, refer to something else than themselves, to the real or imaginary situation that is experienced, believed to exist, conceived or desired. From a cybernetic or "situated" perspective, cognition is necessarily intentional, as mental processes are tighly coupled to, or directed at, external situations via a perception-action feedback loop [Heylighen & Joslyn, 2001]. For example, the internal state of a thermostat, the simplest of cybernetic systems, directly refers to the temperature in the environment, and it is possible to infer the external state from the internal one, and vice-versa.

But as cognition evolves and becomes more sophisticated, the cognitive system becomes increasingly independent of this direct perception-action coupling with the present situation, evolving indirect references to situations that have been, will be, could be or might have been. The reason is simple: the phenomena relevant for the agent's purposes are not always within reach for perception and/or action [Cantwell Smith, 1996], yet it is important to be prepared whenever they might come within reach. For example, it is worth reflecting how you can get protection from the sun even during the night when there is no sun, because of your belief based on experience that the sun will reappear in the morning. Therefore, cognitive systems have evolved the capacity to expect, anticipate, conceive or imagine situations that are not presently available to the senses, and that never may be.

Still, even when a cognitive process refers to an immaterial phenomenon, the very notion of cognition implies some kind of preparedness. This includes at least the capability to recognize the phenomenon if ever it would appear, and to infer some of its properties. To use a classic philosophical example, even though I know that Pegasus does not exist in any physical, material sense, I am prepared to expect a winged, flying horse whenever there is a mention of some Pegasus-like creature, e.g. as depicted in a myth, novel or movie. Thus, Pegasus can have a definite internal reference or "intension" without need for an material, external reference or "extension".

Given this conception of intentionality, how does it apply to distributed cognitive systems as conceived here? At the most abstract level, a collective of agents that has self-organized so as to develop a coherent, adaptive organization is a cybernetic system. As such, it has intentionality: because of the adaptation of the system to its environment, its states will to some degree be correlated with the states of the environment, and thus can be said to be in a relation of reference. Given the on-going learning processes this relation between internal and external states becomes increasingly detailed and subtle, developing an ever more sophisticated relation of anticipation, so that internal states can refer to external situations that have not (as yet) occurred. At the most basic level, this anticipation takes the form of "if...then..." rules connecting learned categories as sketched above, where neither the "if" nor the "then" need to represent actually occurring conditions.

In the simplest case, categories and rules are shared by all agents. In this case the collective intentionality of the system seems little more than an aggregate of individual relations of reference, although as noted above, these individual concepts are likely to have themselves emerged from collective interactions. But when there is cognitive differentiation (i.e. specialization or division of labor) between the agents, there may be little overlap between the concepts and rules used by each individual (Martens, 2004). Still, because of the coordination between the individual information processes, the collective as a whole can be said to have an integrated cognition, distributed over all the agents and their communication media. This implies a non-reducible intentionality from the state of the collective to a perceived or imagined external situation.

For example, during the preparations for the recent US invasion of Iraq, one could say that the American nation had a collective idea of what the Iraqi situation was and what should be done about it. While most individuals shared the basic notion that there was a threat of weapons of mass destruction (WMD) being held by Saddam Hussein, the more detailed perceptions, beliefs and intentions about the situation were distributed among a great diversity of actors, including government, intelligence services, and the military, and coordinated via a dense network of communication channels. No single agent had a complete picture, but the different cognitions were sufficiently coherent and organized to allow the collective to succesfully perform the complex, coordinated action that is a military invasion.

Yet, the immediate success of the action did not entail any "material" correctness of the accompanying beliefs, as later investigation showed that an important part of the initial conceptions that supported this action, such as the assumed presence of WMD, referred to situations that did not exist in reality. Further investigations into the causes for this misjudgment followed the traditional reductionist strategy of looking for the agent (or agents) responsible. While some agents (such as government ministers or intelligence officials) clearly had more influence than others on the final decision, from our perspective it would seem that the incorrect judgment was collective and emergent: the interaction between a variety of agents with different perspectives but linked by relations of trust and mutual support created a distributed dynamic ending up in the attractor of majority consensus.

For example, the publication by the goverment of the imminent WMD threat was based on reports compiled by the intelligence agencies, which themselves were based on a variety of piecemeal observations and interpretations. As the further investigations showed, none of these data in themselves provided convincing evidence, but the recursive process of selecting, compiling, interpreting, reformulating etc. in which one agent built on the assumptions of several other, trusted agents to produce even stronger assumptions which were then fed back to the initial agents to motivate them to come up with further evidence created a self-reinforcing dynamics. This produced a collective conclusion that was much more forceful than what could be warranted by the observations, and that—given the present level of evidence—appears to have been wholly off-the-mark.

Yet, there is nothing exceptional about this particular example of "groupthink" [Janis, 1972]: it merely illustrates the general propensity of self-organizing, connectionist networks to infer clear conclusions from ambiguous data by collating weak, distributed signals from a variety of sources and amplifying them through pre-existing biases and positive feedbacks [McLeod et al., 1998]. As such, it can be viewed as a concrete example of the dynamics underlying the emergence of collective intentionality, whose influence can be summarized by our list of criteria determining which information is preferentially propagated: consistency, simplicity, conformity, authority, utility, etc.

Conclusion

After reviewing diverse perspectives on collective, distributed and extended cognition, we have concluded that although the connections between these approaches are obvious, the domain lacks a unified theoretical framework [cf. Susi & Ziemke, 2001]. Such a framework would have great practical as well as conceptual benefits. For us, the most fundamental issue to be resolved is how distributed cognition can emerge from simple action.

We have argued that a group of interacting agents will tend to self-organize, creating a coordinated division of labor. The resulting social organization will moreover tend to coopt external media for communication. The more a medium is used the more effective it tends to become. Thus, the network formed by the agents connected by their communication links exhibits a form of connectionist learning, characterized by the reinforcement of succesful links, and the weakening of the others.

"Activation" is propagated in parallel along the links, allowing the network to process information in a distributed manner, and fill in missing data. Information is communicated selectively, potentially allowing the essence to be distilled, but also essential diversity lost. The recurrent, non-linear nature of the communication network moreover makes it possible for novel concepts, symbols and intentional relations to emerge in a distributed way. The resulting intentionality can in general not be reduced to the mental states of individual agents but is truly collective.

We are aware that many of these assumptions are quite strong, and may seem exaggerated to some. Yet, we believe that the underlying logic of self-organization gives them a solid foundation. Whether they are also practically useful will have to be ascertained by further research, and in particular by empirical observation and computer simulation of distributed systems.

References

Ashby, W. R. (1962). Principles of the Self-organizing System. In von Foerster, H. and G. W. Zopf, Jr. (Eds.), *Principles of Self-organization*. Pergamon Press, pp. 255-278.

Aunger R. (ed.) (2001): Darwinizing Culture: The Status of Memetics As a Science (Oxford University Press)

Axelrod, R. M., The Evolution of Cooperation, Basic Books New York (1984).

Belpaeme T. (2001) Reaching coherent color 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. and Theraulaz G. (1999) Swarm intelligence: From natural to artificial systems. Oxford University Press.

Brauer, M., Judd, C. M., & Jacquelin (2001). The communication of social stereotypes: The effects of group discussion and information distribution on stereotypic appraisals. *Journal of Personality and Social Psychology, 81*, 463—475.

Cantwell Smith, B. (1996): On the origin of objects (MIT Press)

Chielens, K. (2003) The Viral Aspects of Language: A Quantitative Research of Memetic Selection Criteria. Unpublished Masters Thesis VUB.

Clark A. and Chalmers D. (1998):, "The Extended Mind," Analysis 58, p. 7-19.

Clark, A. (1997). Being There: putting brain, body, and world together again, Cambridge, Mass., MIT Press.

Crowston, K. (2003). A taxonomy of organizational dependencies and coordination mechanisms. In Malone, T. W., Crowston, K. and Herman, G. (Eds.) Tools for

Organizing Business Knowledge: The MIT Process Handbook. Cambridge, MA: MIT Press.

Crutchfield J., Shalizi C., Tumer K & Wolpert D. (eds.) (2002): Collective Cognition Workshop Proceedings: Mathematical Foundations of Distributed Intelligence (http://www.santafe.edu/~dynlearn/colcog/, to be published in the Santa Fe Institute Studies in the Sciences of Complexity, Oxford University Press;)

Crutchfield, J. (1998). Dynamical embodiments of computation in cognitive processes.The Behavior and Brain Sciences, 21, 635

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

Gershenson C. & Heylighen F. (2004): Protocol Requirements for Self-organizing Artifacts: Towards an Ambient Intelligence, in: Proc. Int. Conf. on Complex Systems (New England Institute of Complex Systems)

Hales, D., and B. Edmonds, "Evolving social rationality for MAS using "tags"", Proceedings of the 2nd International Conference on Autonomous Agents and Multiagent Systems (J. S. R. et Al. ed.), ACM Press (2003), 497–503.

Heylighen F. & Bollen J. (2002): "Hebbian Algorithms for a Digital Library Recommendation System", in *Proceedings 2002 International Conference on Parallel Processing Workshops* (IEEE Computer Society Press)

Heylighen F. & C. Gershenson (2003): "The Meaning of Self-organization in Computing", *IEEE Intelligent Systems* 18:4, p. 72-75.

Heylighen F. & Dewaele J-M. (2002): "Variation in the contextuality of language: an empirical measure", *Foundations of Science 6,* p. 293-340

Heylighen F. & Heath M. (eds.) (2004): The Global Brain, special issue of *Technological Forecasting and Social Change* (in press)

Heylighen F. & Joslyn C. (2001): "Cybernetics and Second Order Cybernetics", in: R.A. Meyers (ed.), *Encyclopedia of Physical Science & Technology* (3rd ed.), Vol. 4 , (Academic Press, New York), p. 155-170.

Heylighen F. (1997): "Objective, subjective and intersubjective selectors of knowledge", *Evolution and Cognition* 3:1, p. 63-67.

Heylighen F. (1998): "What makes a meme successful?", in: *Proc. 16th Int. Congress on Cybernetics* (Association Internat. de Cybernétique, Namur), p. 423-418.

Heylighen F. (1999): "Collective Intelligence and its Implementation on the Web: algorithms to develop a collective mental map", *Computational and Mathematical Theory of Organizations* 5(3), p. 253-280.

Heylighen, F. (2003). The Science of Self-organization and Adaptivity, in: Knowledge Management, Organizational Intelligence and Learning, and Complexity, in: *The Encyclopedia of Life Support Systems*, EOLSS Publishers Co. Ltd.

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

Janis, I. L. (1972) *Victims of groupthink.* (Boston: Houghton Mifflin).

Kirsch, D. & Maglio, P. (1994) On distinguishing epistemic from pragmatic action. *Cognitive Science* 18: 513-549

Kirsh, D. Adapting the Environment Instead of Oneself. Adaptive Behavior, Vol 4, No. 3/4, 415-452. (1996)

Klein, O. Jacobs, A., Gemoets, S. Licata, L. & Lambert, S. (2003). Hidden profiles and the consensualization of social stereotypes: how information distribution affects stereotype content and sharedness. *European Journal of Social Psychology, 33*, 755—777.

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

Lyons, A. & Kashima, Y. (2003) How Are Stereotypes Maintained Through Communication? The Influence of Stereotype Sharedness. *Journal of Personality and Social Psychology, 85*, 989-1005.

Martens B. (2004): "The Cognitive Mechanics of Economic Development and Social Change" (PhD thesis, Vrije Universiteit Brussel)

McLeod, P., Plunkett, K. & Rolls, E. T. (1998). *Introduction to connectionist modeling of cognitive processes*. Oxford, UK: Oxford University Press.

Riolo, R., M. D. Cohen, and R. M. Axelrod (2001), "Evolution of cooperation without reciprocity", Nature 414, 441–443.

Searle, J. (1995): The Construction of Social Reality, Free Press

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.

Susi, T. & Ziemke, T. (2001). Social Cognition, Artefacts, and Stigmergy: A Comparative Analysis of Theoretical Frameworks for the Understanding of Artefact-mediated Collaborative Activity. *Cognitive Systems Researc*h, 2(4), 273-290.

Van Overwalle, F., Heylighen F. & Heath M. (2004): From Communication between Individuals to Collective Beliefs (ECCO technical report, http://pcp.vub.ac.be/Papers/SiennaSimulation.pdf).

Van Overwalle, F., & Labiouse, C. (2004) A recurrent connectionist model of person impression formation. *Personality and Social Psychology Review, 8*, 28—61.

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

Wright, R.(2000): *Non-Zero. The Logic of Human Destiny* (Pantheon Books)