A brain in a vat cannot break out: why the singularity must be extended, embedded and embodied

Francis Heylighen
ECCO, Center Leo Apostel,
Vrije Universiteit Brussel

Abstract: The present paper criticizes Chalmers's discussion of the Singularity, viewed as the emergence of a superhuman intelligence via the self-amplifying development of artificial intelligence. The situated and embodied view of cognition rejects the notion that intelligence could arise in a closed "brain-in-a-vat" system, because intelligence is rooted in a high-bandwidth, sensory-motor interaction with the outside world. Instead, it is proposed that superhuman intelligence can emerge only in a distributed fashion, in the form of a self-organizing network of humans, computers, and other technologies: the "Global Brain".

Introduction

This paper is a comment on Chalmers's [2010] discussion of the Singularity, which Chalmers defines as the hypothetical emergence of a superhuman intelligence in the near-term future via the development of an artificial intelligence (AI) system that is so intelligent that it can reprogram itself in such a way as to amplify its intelligence to a level that is higher than the human level (AI+), and eventually so high (AI++) that it can dominate humanity. Chalmers's paper neatly splits into two major issues. First, he discusses the different scenarios for how a Singularity might arise, together with their respective dangers, benefits and ethical implications. In the second part, he discusses how we could mitigate the disadvantages of a Singularity by "uploading", i.e. creating a

digital version of our minds that could integrate with the postulated AI++. Here he again analyses the different possibilities and implications of the idea of transferring a human mind into a digital substrate. Since the paper is much too long to criticize point by point, I will here merely focus on what I consider its major flaw: an implicit ontological assumption of reductionism veering towards dualism. I will demonstrate how this flaw undermines much of the discussion in the first part, and hope that by then it will be clear that the second issue too has become mostly moot.

The paper's title defines it as a "philosophical analysis", thus revealing both the major strengths and weaknesses of Chalmer's treatment of the matter. In terms of strengths, the paper is a very systematic and comprehensive attempt at surveying and clarifying the main conceptual issues surrounding the notion of Singularity. This is particularly useful given that discussions of the Singularity up to now, in the work of authors such as Kurzweil [2005] and Vinge [1993], were extremely speculative, and based on assumptions that, while not a priori implausible, are highly questionable. Chalmers does a thorough job of making those assumptions explicit, formulating equally plausible alternative assumptions, and then asking all the necessary difficult questions. However, apart from some hunches and intuitions, he does not offer much in terms of concrete answers.

This leads us straight into the weaknesses of such an analytic approach: by splitting the problem into a wide array of subproblems, and then splitting up the possible approaches towards solving them into a variety of alternative positions, it seems as if the overall solution only gets further out of reach. The reason is that each subproblem and each position seems to be at least as difficult to tackle as the overall problem: "how can a Singularity emerge?" By way of example, let me mention just one problem that comes out of this analysis, the one that Chalmers in earlier work [Chalmers, 1995] has called the "hard problem of consciousness"—implying that we have at the moment no methodology available to even start investigating it.

This is a deeper issue with all analytic approaches. Analysis by definition means that you divide a whole into parts, hoping that by understanding the parts, you will also understand the whole. While originally formulated by Descartes, this analytic principle is at the basis of classical, Newtonian science, where it led to the philosophy of

reductionism—the idea that you can always reduce the behavior of a whole to the behavior of its parts. Applied to material phenomena, reductionism leads to atomism, the idea that matter is made out of elementary constituents (particles) that completely determine that matter's properties and behavior. While very successful in the realm of physics, such "material" reductionism, however, has largely failed to explain phenomena in the realms of life, mind and society. However, the analytic approach can still be rescued by postulating different, non-material constituents for these phenomena. This is what Descartes did when he observed that mental phenomena did not seem to yield to such a mechanistic reduction: he postulated "mind" as a separate category, independent of matter, so that human behavior could be decomposed into its mental and its physical components. This was the origin of the philosophy of *dualism*, which implicitly seems to inspire much of Chalmers's reasoning, even while he does not explicitly endorse it.

Very few scientists and philosophers nowadays endorse mind-matter dualism. However, many still cling to what Dennett [1993] has called "Cartesian materialism". This can be seen as a "soft" version of dualism, which assumes that mind cannot exist independently of matter, but which otherwise still treats the mind as if it were a separate entity that somehow resides inside our material brain. Chalmers formulates one version of such soft dualism when he characterizes the mind as a brain complemented by unspecified "further facts". It could be argued that traditional cognitive science and artificial intelligence are founded on Cartesian materialism: they see the mind as a piece of software that runs on the hardware of the brain, processing information that enters in the form of symbolic representations of the outside world. This perspective is doubly analytic: it not only separates mind from world, but decomposes information about that world into independently meaningful symbols, each representing a discrete part or aspect of reality. Implicitly, this seems to be the perspective that informs Chalmers and most other Singularity theorists when they discuss the possibility of superhuman artificial intelligence.

The traditional, symbolic approach to AI has undergone increasingly stringent criticisms since the 1980s, from cognitive scientists working from perspectives such as *connectionism* [Bechtel & Abrahamsen, 1991], *constructivism* [Bickhard & Terveen, 1995], *dynamical systems* [Beer, 2000], and especially *situated and embodied cognition*

[Anderson, 2003; Lakoff & Johnson, 1999; Steels & Brooks, 1995; Varela, Thompson & Rosch, 1992]. These "new" approaches to cognition are much more dynamic and holistic, emphasizing the complex network of interactions out of which mind emerges. From this perspective, the mind is no longer localized in any particular component, but distributed over a massive number of internal and external components which all cooperate in a self-organizing manner. Intelligence (and with it consciousness) can then be seen as an emergent phenomenon of coordination between these processes, an integrated manner of dealing with an enormous amount of bits and pieces that together determine an individual's experience of its situation, and that define a potential problem to be dealt with [Heylighen, 2011a, 2007-2011].

From this holistic perspective, trying to understand the mind by analyzing it into components is like eating soup with a knife and fork: no matter how you try to cut up the liquid into pieces, you will never get a piece solid enough so that you can seize it with your fork. Yet, nothing stops you from drinking the soup directly from the bowl, or, if you prefer, spoon by spoon. The trick is simply to use a tool adequate for the task. From my perspective, analytic tools such as logic and symbols appear particularly inadequate for grasping fluid, distributed phenomena, such as intelligence and consciousness. More systemic tools seem much better suited. For example, the "fluidity" of experience can be modeled by activation spreading across a distributed or connectionist network [Heylighen, 2007-2011; Heylighen, Heath & Van Overwalle, 2004].

In the remainder of this paper, I will apply this holistic philosophy of mind, mostly from the perspective of situated and embodied cognition [Anderson, 2003], in order to develop an alternative understanding of the Singularity. I will argue in particular that the "brain-in-a-vat" conception of AI presented by Chalmers is unlikely to lead to any superhuman intelligence, but that such a higher level of cognition (which I call the Global Brain [Heylighen, 2011b]) will necessarily be distributed over a massive number of more or less intelligent components, including individual humans, computers and their programs, databases such as the world-wide web, and a variety of sensors and effectors embedded in the environment. From this point of view, most of the "philosophical" issues raised by Chalmers become irrelevant, while a host of other issues—ethical, political, social and economic—come into focus.

The situated and embodied critique on AI

The view of artificial intelligence sketched by Chalmers and held by most traditional AI researchers has been severely criticized by a variety of authors. These researchers [e.g. Steels & Brooks, 1995; Varela, Thompson & Rosch, 1992; Lakoff & Johnson, 1999; Clark, 2008] come from a variety of backgrounds and points of view, including biology, robotics, linguistics, psychology and philosophy. Yet, they are united by their rejection of what some have called the "brain-in-a-vat" perspective on intelligence. According to this perspective, the only thing you need to produce intelligence is a brain, i.e. a specialized piece of cognitive machinery. You merely need to make sure that the machinery gets the resources (such as matter and energy) that it needs to keep functioning, and the relevant pieces of information to work on. Given those, the brain will process that information and, assuming it is intelligent enough, solve the problems you care to put to it. Whether the "brain" is made out of nerve cells or out of digital circuits is largely irrelevant to that task, as long as those circuits sufficiently accurately simulate the overall functioning of the neurons.

The "situated and embodied" criticism of this view is that the brain has evolved as a specialized organ for the control and coordination of the organism and the actions it performs in its environment. These actions determine whether the organism as a whole will survive and thrive, or fail and eventually be eliminated by natural selection. This entails a constant adaptation to a complex and unpredictable environment, by counteracting or correcting any sensed deviations from a state of fitness. These continuously experienced differences between experienced and desired situations are the "problems" that intelligence was designed to solve [Heylighen, 2007-2011]. In this view, intelligence is not a "thing" or not even a "property", but an on-going process of adaptation and coordination, grounded in a high-bandwidth feedback between organism and environmental situation. Interrupt that interaction and the brain becomes a useless piece of machinery, like a car stationed in an exhibition hall: it may still display some minor features, such as the softness of its chairs or the smooth opening of its electric windows, but it is incapable of doing what it was designed for, driving.

Similarly, a brain in a vat, kept artificially alive via a stream of nutrients, but disconnected from its body, may still be able to produce some abstract thoughts, but it has lost its essential ability to act on the world and to experience the results of its actions. As such, its thoughts will become increasingly disconnected from and irrelevant to reality. Compared to a traditional AI system, such a brain still has an immense advantage, though: it has a lifetime of detailed, subtle experiences stored in its memory on which it can build to perhaps develop new thoughts. Without such history of fine-graded interactions, an AI system is at best a clumsy question-answering "expert system", which needs to be fed a huge amount of symbolic data before it can make any meaningful inference. These data typically need to be structured and formatted by human "knowledge engineers", who abstract and formalize real-world experience into logical expressions. But this is an extremely time-consuming, difficult, unreliable and potentially endless task, known in AI as the "knowledge acquisition" bottleneck. Just throwing in more man-months does not solve the problem, as illustrated by the on-going CYC project [Lenat, 1995], which started out over 25 years ago with the aim of formalizing all common-sense knowledge. The intention was to provide the foundation for an AI-system that would be able to pass something like the Turing test, but as yet without any apparent success.

In the last decade or so, the focus of AI has therefore shifted towards machine learning and data mining, i.e. letting the computer program itself extract knowledge from the huge amounts of data available in specialized databases and on the web. This approach has been much more successful, as recently illustrated by the IBM program Watson, which managed to win a *Jeopardy!* question-answering game against human experts. However, what AI enthusiasts observing this unmistakable advance tend to neglect is that: 1) the knowledge of Watson is not "artificial"; it is the product of millions of humans entering terabytes of text into websites and databases. Without this "extended memory" created by actual agents interacting with the real world, Watson and similar programs would be absolutely helpless; 2) Watson is still nothing more than a passive system waiting until someone asks it a well-formatted question before it can start producing an answer. In the real world, intelligence entails autonomy, i.e. the ability to directly experience the environmental situation and to decide what it means, what

problems may need to be solved, and what actions may be worth taking—without guidance from a programmer, experimenter, or quizmaster.

The newer approaches to AI have started to take the autonomy issue seriously, by creating robots that interact with the world via their sensors and effectors, trying to reach their (preprogrammed) goals, while learning from their experience [Steels & Brooks, 1995]. These autonomous robots at best reach an intelligence level comparable to a primitive insect. But that is not just a question of lack of computational capacity or insufficiently sophisticated software: robots are handicapped by an alternative version of the knowledge acquisition bottleneck. It turns out to be extremely difficult to engineer and build truly efficient systems of sensors and effectors. That is why present-day autonomous robots are very clumsy creatures, who at best may succeed in mowing a flat, well-defined area of lawn, but already start to get in trouble when they have to vacuum a typically more irregular apartment floor.

The awkwardness of engineered sensors and effectors should not surprise us if we remember that evolution needed billions of years to build our extremely sophisticated sensory and motor systems. The complexity resides not only in our eyes, ears, and limbs, but in the integrated system of sensory organs, nerves, hormones, neurotransmitters, brain, glands, muscle fibers, organs, and in fact every cell, which reacts in real time to thousands of chemical signals traveling across the body and which are directly or indirectly triggered by a variety of sensed conditions, such as heat, pressure, smells, stress, emotions, remembrances, etc. For the foreseeable future, this exquisitely coordinated and fine-tuned system seems impossible to reproduce by present engineering techniques, even taking into account "Moore's law" types of acceleration. The reason is that accelerating scientific progress typically happens in well-defined, specialized disciplines, such as chip miniaturization. To build a system rivaling the complexity of the human organism would require massive multidisciplinary coordination and integration of results—the type of progress that still seems as slow and difficult as ever.

The lack of a biological body, therefore, is a fundamental handicap for any artificially intelligent system, even though in theory we can try to build increasingly more sophisticated robotic bodies. But here we are confronted with a hardware bottleneck that

seems much more difficult to overcome than the computational or software bottlenecks that present AI theorists have been focusing on.

Implications for the singularity

Chalmers [2010] envisages the superhuman intelligence emerging from the Singularity as an AI system locked up in some kind of virtual world—a safety measure needed to make sure that this AI++ would not take control of humanity. According to the situated and embodied philosophy of cognition, however, this vision is intrinsically absurd. A virtually imprisoned AI program is even worse than a brain in a vat, as it simply has no sensors, no effectors, no body, and no real world to interact with. Therefore, it cannot be intelligent in the sense of being an autonomous, adaptive cognitive system that can deal with real-world problems and steer its own course of action through a complex and turbulent reality. At best, it can be a very sophisticated expert system that can solve chess, *Jeopardy!*, and similar highly artificial and constrained puzzle and games given to it by its designers, or help them to mine massive amounts of pre-formatted data for hidden statistical patterns. The idea that on the basis of such data it could "reverse engineer human psychology" to such a degree that it could manipulate its creators to give it control over them, as Chalmers proposes, seems more like a paranoid fantasy than like a realistic scenario.

All the arguments put forward by Chalmers to support the AI++ scenario revolve around a runaway process of self-amplifying intelligence. This assumes basically a process of *positive feedback* or *increasing returns*. However, as Chalmers himself notes, positive feedback explosions always come to a halt when the resources for further growth are exhausted, and returns start to diminish rather than increase. This is obvious for processes of physical growth, which are limited by the law of conservation of matter and energy. Things are subtler for cognitive processes, as information does not obey any clear conservation laws: you can destroy or duplicate information freely. However, intelligence is more than multiplication of information: it is extracting value, meaning, knowledge, and eventually wisdom, from data.

From the situated, embodied or enactive perspectives, information is meaningful if you can do something with it, i.e. if it helps you to act towards your goals. Our goals derive from our values, which are themselves the product partly of biological evolution towards survival and reproduction of our genes, partly of cultural instruction towards cooperating fruitfully with society. The traditional AI perspective largely ignores the notion of values, except in the idea that an AI system may need to have some goals or constraints programmed into it (like Asimov's laws of robotics). Chalmers briefly discusses the issue, but seems to ignore that values, like knowledge and intelligence, are the product of millions of years of biological and cultural evolution, and of decades of personal experience with a variety of real-world situations. The idea that something as complex, fuzzy and context-dependent as a system of values could either be introduced by fiat into an AI system or develop autonomously inside its "vat" seems highly unrealistic.

The problem is that without a sophisticated and adapted system of values, an agent will find it very difficult to distinguish what is meaningful or important from what is not. This is a traditional problem in AI (e.g. in the form of the "frame problem"): AI systems typically do not know how to distinguish relevant inferences from trivial ones, unless there is a human guiding them by asking meaningful questions. Autonomous robots have a very simple system of values programmed into them (e.g. whenever your energy level is low, find a plug to recharge your battery; avoid colliding with objects) [Steels & Brooks, 1995]. This helps them to decide about their course of action, but hardly manages to reach the level of complexity of an insect. Without sophisticated criteria for distinguishing what is important from what is frivolous, an AI system can hardly be expected to make intelligent, autonomous plans and decisions, as the AI++ scenario assumes.

Let us look in some more detail at the dynamics of self-amplification. Positive feedback processes such as these are well known in the field of complex dynamic systems (also known as non-linear systems or chaos theory) [Strogatz, 2000]. However, eventually these processes always end up in what is called an *attractor*: a region within the system's state space that the system can enter, but cannot leave [Heylighen, 2001]. Once the attractor has been reached, the system may continue moving around within the

attractor, but it cannot jump out and reach a completely different part of its state space. This means that the system's evolution has essentially reached a stationary state where further innovation is no longer possible. While the system is moving towards its attractor, its "fitness" may be seen to increase, but this increase will slow down as it draws nearer to the attractor's boundary, and comes to a halt inside the attractor. A simple example of this dynamics is a mathematical function (e.g. square root of x) that is iteratively applied to a given starting value (say x = 100, resulting in the sequence: 100, 10, 3.16, 1.77, 1.33, …). In the limit, this recursive application ends in a "fix point" of the function (in this case, x = 1), which is a zero dimensional attractor. Initially, the value of x changes rapidly, but as the self-application is repeated and the fix point comes nearer, further progress becomes infinitesimally small.

The same dynamics can be expected from a hypothetical self-improving AI-system: as it reprograms itself to become more intelligent, the gains in intelligence will become ever smaller, until the process has become to all practical effects stationary. Of course, we do not know how large the initial gains may be, and it is not a priori to be excluded that these gains are large enough to reach AI+ or perhaps even AI++. However, the situated and embodied perspective makes it very unlikely that a mere self-application without environmental interaction will be sufficient to create something that is significantly more intelligent than a system (a human being) that is the product of millions of years of evolution, complemented by decades of personal experience.

For an example of an existing self-application process in computer science, let me refer to the method of supercompilation or metacompilation developed by the cybernetician Valentin Turchin [1986]. A compiler is a program that translates a program written in a high level programming language, such as Java, into a sequence of instructions in machine code ready to be executed by the processor. A good compiler will produce efficient code, which is executed in a small number of steps. A supercompiler is a compiler that is applied to itself to improve its code by bootstrapping. One self-application may make it twice as fast. However, a subsequent self-application may only make it 1.3 as fast, and a third one only 1.1. After a few iterations, further improvements are negligible, and the process reaches its fix point.

Another example relevant to AI is the complexity limit experienced by Artificial Life researchers in their simulations of self-organizing virtual organisms [Rasmussen et al., 2001]. While there exist many successful simulations of evolving ecosystems exhibiting an increasing diversity and complexity of artificial life forms, after a number of iterations complexity always starts to stagnate, and real novelty no longer seems to be created. This should not surprise the reader: if open-ended evolution of virtual life forms was easy to achieve, we would by now already have pretty sophisticated and perhaps even intelligent artificial organisms. Indeed, once a simulation runs successfully, there is no incentive to stop the computer from running ever more iterations, thus producing ever more advanced generations.

Note that this reaching of a stationary or equilibrium state in a closed self-organizing system is also observed in reality for a variety of physical and chemical processes (e.g. crystallization or magnetization) [Heylighen, 2001]. Moreover, it can be justified theoretically, both on the basis of the second law of thermodynamics, and on the basis of the more abstract, functional reasoning of the cybernetician Ashby [1962], who was the first to formulate the concept of self-organization. Open-ended self-organization only occurs in open systems that interact with their environment [e.g. Prigogine & Stengers, 1984]. As complexity scientists have repeatedly pointed out [e.g. Kauffman, 1995], it is the flow of matter, energy and information entering and leaving the system that keeps it dynamic and adaptive—preventing it from reaching a "frozen" equilibrium, fix point or attractor state, while maintaining it on the "edge of chaos" where all the interesting things happen. This fits in perfectly with the situated and embodied theory of cognition, which similarly observes that on-going interaction with the world is necessary to develop and maintain a flexible intelligence.

In conclusion, there are plenty of arguments that make it very implausible that an AI system residing in a closed, virtual world could ever develop a superhuman (AI++) level of intelligence. On the other hand, just giving the AI system some rudimentary sensors and effectors is hardly more likely to bring it to a par with the very sophisticated, billions of years old sensory-motor interface that characterizes the human body. Does this mean that superhuman intelligence remains out of reach? Not at all, as I will try to show in the next section.

The Global Brain

Closely related with the situated and embodied perspective on cognition are the extended mind [Clark, 2008], distributed cognition [Hutchins, 1991; Heylighen et al., 2004] and collective intelligence [Heylighen, 1999] perspectives. Once we have stopped restricting our search for intelligence to a specific, localized component, such as the brain, we become aware that it extends not only across body, sensor, effectors, and the feedback loops between them, but across artifacts used to support cognition (such as signs, notebooks, and computers), across different individuals who help each other solve problems, across organizations, across society, and in a sense across the environment as a whole. All these parts and aspects of our world contribute to our problem-solving and decision-making processes, by providing, storing and processing some of the crucial information.

It is on this level of distributed cognition that we should expect the greatest increases in intelligence. Indeed, the true revolution brought about by accelerating advances in information technology is to be found not in stand-alone AI systems, but in the network of wired and wireless connections, computers, people, organizations, websites, smartphones, embedded chips, sensors, etc. that together form the Net or the Web. This is a truly open, complex, adaptive and interactive system that processes billions of times more information than the most sophisticated stand-alone computing system. Any technological advance—such as faster processors, larger memories or more intelligent programming—that could benefit a stand-alone system will simultaneously benefit the Web. Therefore, there is no reason to assume that the existing gap in capabilities between an individual AI system and the collective system formed by all networked computers and their users will ever decrease.

As I have argued in several papers over the past decade and a half [Heylighen, 1996, 1999, 2007ab, 2011b], there are plenty of observations as well as theoretical arguments for believing that this collective system, which may be called the *Global Brain* [Heylighen, 2011b; Goertzel, 2002], is not only intelligent, but becoming quickly more

intelligent. The reason is that its self-organization is facilitated and accelerated by the seemingly unpreventable processes of globalization and of the increasing spread of information and communication technology. The result of these processes is that information about what is happening on this planet becomes ever more easily available, helping us to make better, more informed decisions, and to tackle more complex problems.

A well-known illustration is how you can find the answer to almost any question within seconds by using the Google search engine. This is not so much because the Google software is particularly smart (although some of the embedded AI definitely helps), but because it makes use of two components of humanity's collective intelligence:

1) with billions of people contributing information and knowledge to the Web (e.g. via the Wikipedia knowledge base, via the millions of blogs and discussion forums and the thousands of newspapers and scientific journals published electronically) for practically any problem for which someone has thought of a good solution, that solution is likely to be available somewhere on the Web;
2) out of the thousands of pages with potential answers to the question, the Google engine can select the ones most likely to be relevant by relying on the (implicit) selections made by millions of users when they clicked on, or made links to, pages they found particularly interesting.

Google's search engine is so successful because it is particularly adept at mining the implicit preferences (i.e. *values*) expressed by people who use the web—thanks in part to its underlying PageRank algorithm [Heylighen, 1999]. As I have argued earlier, the lack of a sense of value, importance or relevance is one of the crucial problems that hamper the development of an autonomous AI system. But any attempt to "hardwire" values into such a system will only provide a very small, restricted and rigid surrogate for the ever-adapting range of values expressed implicitly by the billions of people in the trillions of choices they make every year, while deciding which product to buy, which charity to donate to, which web page to read, or which person to connect to.

Thanks to the most recent technologies—such as smartphones, wireless Internet access, social software, and Twitter—the Global Brain now moreover provides *real-time*

information. This can be used to help individuals on the road (e.g. by guiding them to the most popular nearby restaurant), but also to coordinate complex collective actions (e.g. the protests in the Arab world that toppled several dictatorial regimes). That means that the Global Brain's intelligence is not just a bookish faculty for looking up pre-existing answers in a gigantic library, but a truly interactive ability to navigate a complex and ever-changing environment—for individuals as well as for collectives. The accelerating incorporation of a multitude of hardware sensors and effectors (e.g. satellites, cameras, remote controls) together with an ever more intimate individual-net interface turn the global brain into a truly situated and embodied intelligence. Its powers of interaction with the planetary environment are already orders or magnitude larger than those of the most advanced systems imagined by Singularity theorists. As ICT becomes ever better, and the people learn to use it ever more efficiently, the intelligence of the Global Brain will continue to explode. Arguably, it already has reached a superhuman level. With a collective intelligence that powerful, who needs a stand-alone AI++?

My point is that it is simply much easier, cheaper and more effective to augment intelligence by facilitating distributed cognition than by building localized, autonomous AI systems. The collective intelligence of the Global Brain communicates with the world via a channel whose bandwidth is many orders of magnitude larger than even the most powerful imaginable stand-alone system. Every human being or piece of machinery hooked up to the web forms part of its body, providing it with additional capabilities for input, output and control. The only thing needed to use this power effectively is *coordination*. As I have argued in several publications [e.g. Heylighen, 2001, 2007a, 2011a], such coordination self-organizes easily once the main sources of friction (such as time, distance, and requirement for energy) are removed—something the Internet has to an important degree already achieved. As systems are standardized (e.g. via the Semantic Web initiative), as new tools for coordination are developed (e.g. social networks, wikis), and as people learn how to use these tools more effectively, we see this self-organization of collective cognition take place at an absolutely staggering rate. Extrapolating this accelerating advance leads us to expect for the near-term future a true technological "Singularity"—in its original mathematical sense as a discontinuous transition beyond

which further extrapolation becomes impossible [Heylighen, 2007a], not in the more limited sense used by Chalmers [2010] as a superintelligent AI system.

In contrast, when we look at the development of stand-alone AI over the last half century we see little or no spectacular progress: we just see expert systems becoming somewhat better experts at the highly specialized tasks (such as playing chess or parsing language) that they have been programmed to perform. The reason for the different rates of progress is that the Global Brain is—by its very nature—extended, embodied and embedded into the real world, with all its complex and ever-changing ramifications. This allows it to soak in information and to directly intervene via billions of high-bandwidth interfaces. The feedback flows between these inputs and outputs drive its self-organization towards ever more complex and adaptive intelligence. Stand-alone AI, on the other hand, is a poor analogue of a brain in vat, which must be artificially fed with preformatted knowledge and told what to do by its programmers. Self-application, as envisaged by Singularity thinkers, may help it to bootstrap some additional capacity, but is unlikely to give it anything that the Global Brain could not develop at least as well.

Conclusion

I have given a variety of arguments for why Chalmers's [2010] conception of the emergence of a superhuman artificial intelligence within a virtual world is very unlikely to happen, at least if this intelligence is supposed to have sufficient common sense and real-world understanding to be able to intervene in and potentially control humanity's affairs. The reason is that the defining character of such an AI++—its independent, closed, self-creation—is antithetical to the now dominant understanding of cognition as a process that is necessarily extended, embodied and embedded into the outside world.

However, this does not mean that I reject the possibility of the near-term emergence of a superhuman intelligence, i.e. a Singularity. On the contrary, the perspective of distributed cognition makes such a "metasystem transition" [Turchin, 1977; Heylighen, 2007ab] to a higher level of intelligence and organization rather more likely. Social systems, supported by a variety of tools for storing, processing and

propagating information, have always exhibited some degree of collective intelligence above and beyond the intelligence of their individual components [Hutchins, 1991]. The tremendous advances in information technology that Singularity theorists like to contemplate have had their biggest impact on this collective, distributed level—not on the level of stand-alone computing systems.

An anecdote to illustrate this observation can be found in the birth around 1970 of the computer network, which was originally designed to allow researchers to perform calculations on a mainframe computer in a far-away location. When a researcher would use the network connecting these huge machines to leave a message for another researcher working at a different institute, this seemed merely a quick and dirty hack. Yet, this "improper" use of the network for exchanging information (which was later to evolve into email and eventually the world-wide web) soon became its most popular application, while its original function of logging in to a remote computer has virtually disappeared by now.

It could be argued that Singularity theorists like Chalmers are still in mainframe mode, standing in awe for the tremendous power of those huge, stand-alone computers that can perform calculations faster than any human being. I see computers in their email and web modes, as fast and efficient intermediaries that process and propagate information across a hugely complex network of people and things, but that have essentially disappeared into the background. It is in the self-organization of this immense network that a superhuman level of intelligence is most likely to emerge. This intelligence will not be localized in any particular component, but distributed over all its components, human and machine. The human components are essential for the efficient functioning of this intelligence, because they are the only ones that, at least for the foreseeable future, offer a true embodiment, i.e. the ability for high-bandwidth interaction with the world via sophisticated sensors and effectors.

In the longer term, as hardware abilities expand, the need for individual humans may diminish, but this is unlikely to follow a simple "takeover by the robots" scenario. If we view the network as a superorganism with both hardware and organic components that become ever more intimately linked (e.g. via brain-computer interfaces), then the transition from a biology-dominated to a technology-dominated Global Brain is likely to

be a smooth, extended process with no clear switch from one regime to a another. In the process, an increasing amount of human experience is likely to become "uploaded" to a digital medium, thus blurring the boundaries between human and technological systems. This raises a lot of philosophical, ethical but especially practical, psychological and social questions, which to some degree echo Chalmers's concerns. However, I hope I have made it clear that the theories of embodied and distributed cognition formulate the issue in a fundamentally different way, making most of the questions raised by Chalmers moot, while raising a bunch of different questions. For a first exploration of these questions and some preliminary answers, I refer to earlier publications [Heylighen, 2007ab; Heylighen & Goertzel, 2011].

References

Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial intelligence*, *149*(1), 91–130.

Ashby, W. R. (1962). Principles of the self-organizing system. *Principles of Self-Organization*, 255–278.

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the mind: An introduction to parallel processing in networks*. Basil Blackwell.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in cognitive sciences*, *4*(3), 91–99.

Bickhard, M. H., & Terveen, L. (1995). *Foundational issues in artificial intelligence and cognitive science: Impasse and solution*. North Holland.

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219.

Chalmers, D. (2010). The Singularity: A philosophical analysis. *Journal of Consciousness Studies, 17*, *9*(10), 7–65.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford University Press.

Dennett, D. C. (1993). *Consciousness explained*. Penguin London.

Goertzel, B. (2002). *Creating internet intelligence: Wild computing, distributed digital consciousness, and the emerging global brain*. Kluwer Academic/Plenum Publishers.

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

Heylighen F. (2001): "The Science of Self-organization and Adaptivity", in: L. D. Kiel, (ed.) Knowledge Management, Organizational Intelligence and Learning, and Complexity, in: The Encyclopedia of Life Support Systems ((EOLSS), (Eolss Publishers, Oxford).

Heylighen F. (2007-2011) Cognitive Systems: a cybernetic perspective on the new science of the mind (ECCO working paper 2007-07), lecture notes for the course "Cognitieve Systemen", available at http://pcp.vub.ac.be/Papers/CognitiveSystems.pdf

Heylighen F. (2007a): Accelerating Socio-Technological Evolution: from ephemeralization and stigmergy to the global brain, in: "Globalization as an Evolutionary Process: Modeling Global Change", edited by George Modelski, Tessaleno Devezas, and William Thompson, London: Routledge, p.286-335.

Heylighen F. (2007b): "The Global Superorganism: an evolutionary-cybernetic model of the emerging network society", Social Evolution & History. 6 No. 1,p. 58-119

Heylighen F. (2011a) Self-organization of complex, intelligent systems: an action ontology for transdisciplinary integration, Integral Review (in press)

Heylighen F. (2011b): "Conceptions of a Global Brain: an historical review", in: Evolution: Cosmic, Biological, and Social, pp: 274 - 289, eds: Grinin, L. E., Carneiro, R. L., Korotayev A. V., Spier F., Uchitel Publishing, Moscow.

Heylighen F. & Bollen J. (1996) "The World-Wide Web as a Super-Brain: from metaphor to model", in: Cybernetics and Systems '96 R. Trappl (ed.), (Austrian Society for Cybernetics).p. 917-922.

Heylighen F. & Goertzel B. (2011): Francis Heylighen on the Emerging Global Brain: An Interview by Ben Goertzel, *h+ Magazine* (March 16, 2011), http://hplusmagazine.com/2011/03/16/francis-heylighen-on-the-emerging-global-brain/

Heylighen F., Heath M., F. Van Overwalle (2004): The Emergence of Distributed Cognition: a conceptual framework, *Proceedings of Collective Intentionality IV*, Siena (Italy)

Hutchins, E. (1991). Social organization of distributed cognition. In L. Resnick, J. Levine, & S. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 283–307). Washington, DC: American Psychological Association.

Kauffman, S. A. (1995). *At home in the universe*. Oxford University Press New York.

Kurzweil, R. (2005). *The singularity is near*. Penguin books.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. Basic books.

Lenat, D.B., (1995). CYC: A large-scale investment in knowledge infrastructure, *Communications of the ACM* 38 , Issue 11, p. 33 – 38.

Prigogine, I., Stengers, I. (1984). Order out of chaos. *Bantam Books,*

Rasmussen, S., Baas, N. A., Mayer, B., Nilsson, M., & Olesen, M. W. (2001). Ansatz for dynamical hierarchies. *Artificial Life*, *7*(4), 329–353.

Steels, L., & Brooks, R. A. (1995). *The artificial life route to artificial intelligence: Building embodied, situated agents*. Lawrence Erlbaum.

Strogatz, S. H. (2000). Nonlinear dynamics and chaos. Westview Press.

Turchin, V. (1977). The Phenomenon of Science. A Cybernetic Approach to Human Evolution. *New York: Columbia University*.

Turchin, V. F. (1986). The concept of a supercompiler. *ACM Transactions on Programming Languages and Systems*, *8*(3), 292-325.

Varela, F. J., Thompson, E., & Rosch, E. (1992). *The embodied mind: Cognitive science and human experience*. The MIT Press.

Vinge, V. (1993). The coming technological singularity. *Whole Earth Review*, 88–95.