Human-Level Artificial General Intelligence and the Possibility of a Technological Singularity

A Reaction to Ray Kurzweil's *The Singularity Is Near*,
and McDermott's Critique of Kurzweil

Ben Goertzel

1 Narrow AI versus AGI

The AI field started out with grand dreams of human-level artificial general intelligence. During the last half-century, enthusiasm for these grand AI dreams – both within the AI profession and in society at large -- has risen and fallen repeatedly, each time with a similar pattern of high hopes and media hype followed by overall disappointment. Throughout these fluctuations, though, research and development have steadily advanced on various fronts within AI and allied disciplines.

Averaging across the various historical fluctuations, we may generalize that the original vision of human-level AI has been dampened over time due to various coupled factors, including most prominently

- overoptimistic promises by early AI researchers, followed by failures to deliver on these promises (Crevier, 1993; Dreyfus, 1992)
- a deeper understanding of the underlying computational and conceptual difficulties involved in various mental operations that humans, in everyday life, consider trivial and simple (Minsky, 2006; Thagard, 2005)

These days most R&D carrying the label "AI" pertains to some sort of very narrowly-defined problem domain, shying away from the ambitious goals that are associated with AI in the popular media.

Recently, in the first years of the 21st century, AI optimism has been on the rise again, both within the AI field and in the science and technology community as a whole. One possibility is that this is just another fluctuation – another instance of excessive enthusiasm and hype to be followed by another round of inevitable disappointment. Another possibility is that AI's time is finally near, and what we are seeing now is the early glimmerings of a rapid growth phase in AI R&D, such as has not been seen in the field's history to date.

As evidence of the recent increase in AI optimism in some relevant circles, I note that the last few years have seen an increasing number of conference special sessions and workshops focused on "Human-Level AI," "Artificial General Intelligence" and related topics; for example (this is not a comprehensive list):

- *Integrated Intelligent Capabilities*, Special Track of AAAI since 2006 (and planned to continue annually into the indefinite future)
- *Roadmap to Human-Level Intelligence*. Special Session at WCCI, July-2006
- *Building & Evaluating Models of Human-Level Intelligence*, CogSci, July-2006
- *Artificial General Intelligence Workshop*, Bethesda MD, May-2006
- *Between a Rock and a Hard Place: Cognitive Science Principles Meet AI-Hard Problems*, AAAI Spring Symposium Series, Mar-2006
- *Towards Human-Level AI*, NIPS Workshop, Dec-2005
- *Achieving Human-Level Intelligence through Integrated Systems and Research*, AAAI Fall Symposium Series, Oct-2004

In March 2008, the University of Memphis will host AGI-08, the first AI conference focusing specifically on artificial general intelligence and human-level AI, co-organized by the author and several colleagues, including Stan Franklin, a long-established leader in the AI field. There have also been specific attempts to focus publications on human-level artificial general intelligence, e.g. a "Human-Level AI" issue of AI Magazine edited by Nick Cassimatis (2006), and a series of edited volumes focused on AGI (Goertzel and Pennachin, 2006; Goertzel and Wang, 2007).

And in the popular press, futurist pundits are more and more often heard proclaiming that the age of real AI is, this time, *really* coming soon – not like the false hopes of the past. The most widely-read example of this sort of futurist AI optimism is probably Ray Kurzweil's recent book *The Singularity Is Near* (*TSIN*; 2005), which projects the achievement of human-level AI by roughly 2029, followed by an AI-triggered radical transformation of mind, society and economy by roughly 2045.

Of course, the vast majority of academic and industry AI researchers remain deeply skeptical of the sort of optimistic rhetoric and perspective that Kurzweil's book typifies. The annual AAAI conference remains focused on the solution of important, fascinating, but very narrowly-defined technical problems, most of which are only loosely (if at all) connected to the problem of creating artificial general intelligence at the human level or beyond. More ambitious AI ideas remain at the periphery.

Kurzweil has called the currently mainstream sort of AI research "narrow AI" – meaning AI that focuses on the creation of software solving specific, narrowly constrained problems. A narrow AI program need not understand itself or what it is doing, and it need not be able to generalize what it has learned beyond its narrowly constrained problem domain. For example, a narrow-AI program for playing chess need not be able to transfer any of its strategic or methodological insights to Shogi (Japanese chess) or checkers ... and probably not even to Fisher random chess (though a human programmer might be able to take some of the knowledge implicit in a narrow-AI chess playing program and use this to make a better program for playing other games; in this case the general intelligence exists mainly in the human being not the programs). A narrow-AI program for driving a car in the desert need not be able to utilize its knowledge to drive a car in the city or a motorcycle in the desert. A narrow-AI program for parsing English cannot learn any other language, whether or not the other language has a similar syntactic and semantic structure. A narrow-AI program for diagnosing kidney cancer will always be useless for diagnosing gall bladder cancer (though the same

narrow-AI framework may be used by humans to create narrow-AI programs for diagnosing various sorts of cancers).

Kurzweil contrasts narrow AI with "strong AI," but I find this terminology confusing due to its overlap with Searle's better-known usage of the term "strong AI," so I prefer the term AGI or "Artificial General Intelligence" for the opposite of Kurzweil's "narrow AI." A related term sometimes used is "Human-Level AI," but I consider this term less preferable for two reasons:

- In the space of all possible minds, humans are not necessarily all that smart, so that the "human level" constraint actually may pose an overly strict limitation on the scope of future AGI work
- Defining what "human level" means is actually difficult, when one starts thinking about potential highly-general-intelligence AI systems with fundamentally non-human-like architectures. If one has an AGI system with very different strengths and weaknesses than humans, but still with the power to solve complex problems across a variety of domains and transfer knowledge flexibly between these domains, it may be hard to meaningfully define whether this system is "human-level" or not.

For the rest of this essay I will stick with the term AGI, though using "Human-Level AI" when that is specifically what I mean.

Now, one might argue that the position of AGI R&D on the margins of contemporary AI research is only correct and proper, since we don't really know how to do human-level AGI yet; and the bulk of contemporary AI research focuses on more narrowly-defined research directions that have the benefit of far more easily leading to scientifically demonstrable and/or pragmatically useful results. However, there are other branches of contemporary science where the overall modus operandi is not so conservative. For instance, physics currently devotes a significant amount of attention to speculative mathematical theories of unified physics, which are no more soundly proven than anybody's current approach to AGI. This work is considered justified because it is advancing understanding, and seems likely (according to the very theories being developed, which are not yet proven) to yield exciting experimental results in future. And, the biopharmaceutical industry has devoted a huge amount of funding to areas such as gene-therapy based medicine, which has not yet led to any dramatic practical successes, and still involves a huge amount of uncertainty (for instance, how to effectively deliver modified genes to the appropriate part of the organism being healed). Quantum computing is the focus of much excitement in spite of the fact that all known quantum computers are extremely specialized or extremely small (a handful of qubits), and the number of fundamental quantum computing algorithms known can be counted on the fingers of one hand. Modern science, in other areas, is willing to take a medium and long term view and focus significant amounts of attention on the big problems. Other examples abound. But the field of AI, due to its particular history, has settled into a rather conservative pattern. Which makes the contrast between the goings-on at the average AAAI conference session, and the prognostications of a futurist AI pundit like Kurzweil, particularly dramatic, poignant and amusing.

My own view is that AGI should be the focus of a significant percentage of contemporary AI research; and that dramatic progress on AGI in the near future is something that's reasonably likely (though by no means certain) to happen. In the remainder of this essay I'll present this view mainly via reviewing some of Ray Kurzweil's recent arguments, and their attempted rebuttal in a recent article by Drew McDermott. I will reframe Kurzweil's ideas and predictions within a larger perspective of scenario analysis, review their strengths and weaknesses in this context, and then consider them side-by-side with other scenarios that are also plausible and in my view at least equally exciting. Among the many important issues to be considered along the way is the nature of the "Singularity" event that Kurzweil refers to in the title of his book, and the relationship between AGI and the Singularity. I realize these issues are fairly far "out there" compared to what most contemporary AI researchers think about from day to day – but I contend that this is something that should, and will, change.

2 Scenario Analysis and the Future of AGI

The future of AGI is a big, hard, complicated issue. No one can sensibly claim to know what's going to happen. Dealing with issues like this is difficult; there is uncertainty that is irreducible (at least from the perspective of mere human-level minds!).

One methodology that has been crafted for dealing with very difficult problems involving complex systems and high levels of uncertainty is "scenario analysis." Originally crafted by Shell Oil planning pundits in the 1970's, scenario analysis has been expanded into a more general methodology, and has been used profitably to deal with a variety of critical real-world situations, such as the establishment of a new order in South Africa following the abolition of apartheid (Kahane, 2007).

The basic idea of scenario analysis is simple. Rather than trying to make definite predictions with specific probabilities, or trying to arrive at elegant abstractions binding the past, present and future, one tries to lay out a series of specific future scenarios for a complex system. Ideally each scenario should be fleshed out by a group of individuals with expertise and intuition regarding different aspects of the system in question, and different opinions on key, relevant controversial issues. Applying scenario analysis to the South African political situation at the time of the abolition of apartheid, for example, involved organizing meetings involving a wide variety of parties, including black anti-apartheid activists, government and business leaders, and so forth. Rather than arguing about what will or won't happen, the goal of the team is to use their collective intuition, knowledge and expertise to flesh out a variety of particular possible scenarios. Once this is done, then analytic and evaluative effort can be expended assessing the plausibility and desirability of various scenarios.

I think this would be an excellent way to confront the question of the future of AGI and its relationship to the rest of technology and society. In the ideal approach, one would convene a series of meetings involving a variety of parties involving AGI researchers, narrow AI researchers, technology pundits, social and business leaders, artists and psychologists, and so forth. The goal would be to explore in detail a variety of plausible scenarios regarding what the future of AGI may hold, both scientifically and in terms of broader implications.

Lacking such an explicit scenario analysis approach to exploring AGI, however, I think scenario analysis can also be fruitfully deployed as a way of structuring and evaluating the thinking and speculating about AGI and its future that's already been done by various knowledgeable individuals and groups. Example of plausible categories of scenarios regarding the future of AGI would be the following (this is certainly not a complete list!):

Steady Incremental Progress Scenarios: Narrow-AI research continues incrementally, as it's doing now – gradually and slowly becoming less and less narrow. Explicit AGI research doesn't really get anywhere till narrow AI has built up a lot more knowledge about how to solve particular sorts of problems using specialized algorithms. Eventually, maybe hundreds of years from now, narrow AI research becomes sufficiently general that it reaches human-level AGI. By that point, AI tech at various levels of generality is already thoroughly integrated into human society, and potentially even into human organisms via brain-computer interfacing technology (as reviewed by Kurzweil in *TSIN*).

Dead-End Scenarios: Narrow-AI research continues and leads to various domain-specific successes but doesn't succeed in progressively moving toward AGI. Explicit AGI research also fails to progress. The human mind proves incapable of understanding, replicating or improving on human-level intelligence.

Each of two these scenario-categories branches into multiple, radically variant scenarios for the future as a whole, depending on one's assumptions about the advances of other futuristic technologies besides AGI, such as neuromodification and nanotechnology (extensively discussed by Kurzweil in *TSIN*; for the classic discussion of nanotech see Drexler, 1992). For instance, if the creation of something vaguely like a Drexlerian molecular assembler becomes possible within the next couple centuries, then the above scenario-categories correspond to scenarios in which human life changes dramatically and unpredictably, but AGI doesn't play a major role in it. On the other hand, if molecular assemblers and other hoped-for radical technologies prove just as difficult as AGI, then these scenario-categories correspond to overall scenarios of incremental social, technological, political and psychological evolution (scenarios which, as it happens, are relatively straightforward to explore, although clearly we have a pretty weak track record at predicting the evolution of human systems even under these relatively "easy" conditions).

On the other hand, a more exciting category of scenarios are the

AGI-Based Singularity Scenarios: A human-level AGI is achieved, and this AGI succeeds at both progressively increasing its own intelligence and creating various other radical technologies, thus leading to a massive and massively unpredictable transformation of the conditions in our region of the universe.

The term "Singularity" was introduced in this context by Vernor Vinge (1993), who used it to refer to the point at which scientific and technological progress occur so fast that, from the perspective of human cognition and perception, the rate of advancement is

effectively infinite. The knowledge-advancement curve becomes effectively vertical, from a human perspective.

The term Singularity in this context also gains some semantic flavor by analogy to singularities in mathematical physics, particularly black holes. This analogy invokes the concept of the "event horizon" which shields the singularity from outside observation. This point of view is informed by the notion that it is difficult for humans to predict what will be done by intellects which exceed our own.

Slightly less dramatically, J. Storrs Hall (2007) has referred to the possibility of a coming "Intellectual Revolution," loosely analogous to the Industrial Revolution. But whether we think about a revolution or a Singularity, many features of the situation remain the same. There is still a dramatic, irreducible uncertainty attached to the development of any future technology as radical as human-level AGI. Whether the change is conceived as a Sngularity or a mere revolution, the character of the human condition following such a major change is substantially unknown, and the outcome may well depend critically on the nature of the AGI programs that enable it.

Within the category of Singularity Scenarios, there are, again, various particular scenarios, differentiated by multiple factors including the manner in which AGI is achieved. As examples we may look at

Skynet Scenario: Named after the malevolent AGI in the Terminator movies. A powerful AGI is created, improves itself, develops amazing new technologies (backwards time travel, in the movies), and enslaves or annihilates us pesky little humans. In the modern futurist literature there is recurrent talk of superhuman AI's transforming all the molecules of their region of the universe into "computronium" – i.e. into processors and memory intended to enhance their own intelligence. After all, from a superhuman AI's point of view, isn't a superhuman intelligence a better use of available mass-energy resources than a bunch of stupid little atavistic humans?

Kurzweil Scenario: The scenario Kurzweil envisions in *TSIN* is essentially that AGI is achieved via scanning human brains, figuring out the nature of human thought from these scans, and then replicating human brain function on massively powerful computer hardware (whose existence is predicted by Moore's Law and its brethren). Furthermore, he views this human-like human-level AGI as integrating richly with human society, so that it's not an "Us versus Them" type scenario, but rather a scenario in which the boundary between Us and Them is fluid, evolving and impossible to define.

While Kurzweil is enthused about human-brain emulation (HBE), this is obviously not the only possible path to AGI, nor will it necessarily be the first path to succeed. Furthermore, once HBE is achieved, this may relatively quickly lead to other forms of advanced AGI (a point I'll elaborate on below).

In *The Path to Posthumanity* (Goertzel and Bugaj, 2006), the author and Stephan Vladimir Bugaj suggest that the most likely future is one in which human-level AGI is

first achieved via integrative methods (synthesizing insights from computer science, cognitive science and other disciplines) rather than via emulation of the human brain. In this vein, a number of contemporary AGI researchers are pursuing computer-science-centric approaches, inspired only loosely by cognitive neuroscience (and generally more so by cognitive psychology proper; e.g Franklin's (2007) LIDA system which is based closely on Bernard Baars' work and SOAR (Jones and Wray, 2004) and ACT-R (Anderson, 2000) which follow various psychological studies very rigorously.) From a Singularitarian perspective, non-human-emulating AGI architectures may potentially have significant advantages over human-emulating ones, in areas such as robust, flexible self-modifiability, and the possession of a rational normative goal system that is engineered to persist throughout successive phases of radical growth, development and self-modification. This leads to a variety of possible futurological scenarios, including three which *The Path to Posthumanity* identifies as particularly interesting:

Sysop Scenario. Yudkowsky (2002) described the "Sysop Scenario" – a scenario in which a highly powerful AI effectively becomes a "system operator" for this region of the physical universe. Goertzel and Bugaj, tongue-in-cheek, referred to a closely related "AI Buddha" scenario, focusing on the Sysop's need to display a practical compassion for humans and other beings who are far its inferiors. The basic feature is the existence of an AGI with dramatically superhuman powers, and a will and mandate to make the universe (or its region thereof) a hospitable place for the various other minds and life-forms resident within it, as well as advancing its own self in directions it finds desirable. This is the ultimate extension of the archetype of the "benevolent dictator."

AI Big Brother Scenario. The task of creating an AGI that will evolve into a Sysop or something else reasonably beneficent may prove an arduous one, perhaps even intractable, which brings up the possible desirability of preventing the advent of super-advanced AI technologies. One way to do this would be to freeze technological development at its current levels; but this seems difficult to do, consistent with contemporary ideals of capitalism and personal freedom. If these ideals are sacrificed, on approach that might be effective would be to create an AGI with greater-than-human capabilities, but without a mandate for progressive self-improvement; rather, with a mandate for preserving the human status quo, using other technology such as nanotechnology (XX) as appropriate.

Singularity Steward Scenario. Finally, there is a variant of the above two scenarios in which a powerful AGI is created with a temporary goal of easing the human race through its Singularity in a smooth and beneficial way. This is the sort of future alluded to in Damien Broderick's novel Transcension (2003).

Finally, another solution, fascinating though fanciful, is outlined in Yudkowsky (2004),

Coherent Extrapolated Volition Scenario. In this scenario, the human race realizes that choosing the right future scenario is too hard, so it creates a specialized narrow-AI optimization process whose goal is to figure out what humans really want, and issue a report (which we then are free to put in a file cabinet and ignore, of course!).

Obviously, the list of future scenarios given above is far from complete. However, I think it is long enough to give a concrete flavor of the raw indeterminacy associated with the possibility of the technological Singularity. As Vinge pointed out when he first articulated the Singularity idea, the key point about the creation of superhuman intelligence is that, once it's here, in all probability we measly humans simply have no way to predict what happens next.

3 Kurzweil's Singularity Scenario and McDermott's Critique

Getting back to Kurzweil, then: One thing that he has accomplished in *TSIN* and related writings is to articulate, in impressive detail and with admirable articulacy, one possible Singularity scenario (the one I have labeled the "Kurzweil scenario" above). In this vein he does three things, all of them well:

1. He describes what this scenario might be like for those living through it
2. He describes the scientific path by which it might come about
3. He gives his own estimates of the probability (very high) and the expected timeline (human-level AI by 2029, Singularity by 2045).

Much of his effort seems to have gone into the detailed graphs and charts supporting the timeline estimates (item 3); as I'll elaborate below, I think these extrapolations are worthy as well as attention-grabbing, but would be more interesting if they came with equally carefully estimated confidence intervals. On the other hand, I think he has done a tremendous service by carrying out item 2 in the above list, and describing, for the general reader, the scientific path by which a Kurzweilian Singularity might pragmatically be brought to pass.

In 2006, this journal published a review by Drew McDermott entitled "Kurzweil's argument for the success of AI", which presents a fairly harsh critique of contemporary AI optimism, focusing heavily on *The Singularity Is Near*. While McDermott's critique does contain some cogent points, it also seems to miss some of the key points Kurzweil's book tries to get across. It seems worthwhile to briefly pause here to consider some of McDermott's counter-arguments and their strengths and shortcomings.

One of McDermott's main points is that Kurzweil does not provide any *proof* that an AI-driven Singularity is upon us. This is certainly the case. However, nor does Kurzweil claim to provide proof – he strives only to provide a convincing rational argument of likelihood. Obviously, in any extrapolation of the future of a complex real-world system coupled with other complex real-world systems, there can be no such thing as proof, only at best "probably approximately correct" prediction.

I do tend to feel that Kurzweil underestimates the uncertainty involved in predicting the future of complex, open systems like human societies (and the scientific and technological development taking place within them). The tendency of the human mind toward overconfidence is well-researched and well-documented in the cognitive

psychology community (Gilovich et al, 2002), and in reading Kurzweil's book, it's hard to escape the feeling that in spite of the rigor of his data analysis procedures, he has succumbed to this tendency to a certain extent. Where are the confidence intervals around his prognostications? Sure, it's useful to have a specific date in mind, to make a prediction concrete and palpable – but how certain could a date like "2045 for human-level AI" possibly be? Even if 2045 is a good estimate of the mean of the distribution of dates for human-level AI, what's the variance? Clearly it must be pretty large – because there are plausible scenarios in which human-level AI is delayed beyond 2100, and also (a more controversial statement, I realize, but one that I'll stand by) plausible scenarios in which it's achieved by 2020 or even 2015. Even "human-level AI by 2010" has a probability non-negligibly greater than zero, I would venture. There is a lot more uncertainty in the future than Ray Kurzweil wants to recognize.

Taking a scenario-analysis point of view, however, gives a different perspective on Kurzweil's achievement as a futurologist. What he has done is to give a very clear portrayal of a particular scenario for our future, and then to argue why this scenario is in fact a reasonably likely one. The fact that he may display a bit of overconfidence in estimating the probability of this scenario over other ones, or the date of the arrival of particular events within this scenario, doesn't detract from the crux of his achievement.

The most interesting point in McDermott's critique is his attempted rebuttal of Kurzweil's argument as to why human-level AGI is likely to occur by 2029. To get this figure, Kurzweil extrapolates not from contemporary progress in the AI field, but rather from contemporary progress in computer hardware and brain scanning. He argues that by 2029 we will have computers powerful enough to host a detailed emulation of the human brain, and brain-scanners powerful enough to mine the brain-data needed to create such an emulation. So according to this argument, even if the AI approaches currently being pursued are all wrong, we'll still get to human-level AI soon enough just by copying the brain.

The most important thing to understand about this argument is that Kurzweil intends it centrally as a sort of "existence proof." He's not saying that human-brain emulation is the only route to human-level AGI, he's mostly just saying that it's a plausible route – and a route that one can argue for by extrapolating historical growth curves of various non-AI technologies. My own view is that Kurzweil is essentially correct in his extrapolations, and that brain-scanning and computer-hardware will conspire to allow effective human brain emulation sometime in the next few decades. Just recently, the newspapers contained reports that IBM had "simulated half a mouse brain" (BBC News, 2007). It turns out that the simulation ran only briefly, ran at 1/10 the speed of an actual mouse brain – and, most limitingly, was in fact just a simulation of the *same number of neurons believed to comprise half a mouse brain*, but with largely random interconnectivities (because brain scanning doesn't yet tell us how a mouse's neurons are interconnected). But still, this sort of development is highly consistent with Kurzweil's projections.

My own guess happens to be that we will achieve human-level AGI via other means, well before the brain-scanning route gets us there. But, this is not an argument against Kurzweil's projections. I think that computer hardware and brain-scanning are simply more predictable technologies than non-human-brain-based AI software. The latter, I suggest, is more likely to experience a sudden and generally-surprising

revolutionary advance; but also more likely to get stuck on unforeseen obstacles and drag on a long time with disappointing progress.

The main dispute McDermott has with the brain-scanning route to AGI is that, even if we succeed in scanning the brain into a computer, this still won't give us any real understanding of how intelligence (human or otherwise) works. In his words,

Obviously improvements in brain-scanning technology are important and exciting, but they get us somewhere only if accompanied by deepening of our understanding of the computations neurons do. So the possibility of scanning is a very weak argument that our understanding is sure to deepen.

...

[E]ven if we succeeded in duplicating a person to the point where we couldn't tell the copy from the original, that wouldn't even confirm AI, let alone contribute to it. (p.6)

More technically, McDermott claims that simulating a brain in computer hardware wouldn't suffice as a demonstration of "computationalism," which he defines as the assertion that

a) there is such a thing as the 'computational state' of a creature, which can be characterized independently of the physical substrate used to implement it'
b) what computational state a system is in complete determines its mental state

And relatedly, he takes issue with

the idea that once we achieve 'strong AI' we will 'soar' on to what I will call ***superstrong AI****. Once more, I don't think the case is proven.*

arguing against Kurzweil's vision of recursively self-improving AI via accusing it of embodying circular reasoning:

" 'Machine intelligence will improve its own abilities in a feedback cycle that unaided human intelligence will not be able to follow.' The last argument is, alas, circular, as far as I can see. Until machine intelligence gets past the human level, it won't be able to do all that smart stuff. "

These objections do, I think, have some point to them. But the point was already understood and addressed by Kurzweil in his book. Note that Kurzweil foresees human-level AGI via brain emulation in 2029, and Singularity in 2045. This is in part because Kurzweil understands that being able to emulate a brain in computer hardware is different from having a real understanding of how to flexibly create human-level artificial intelligences. And Kurzweil understands that the human brain architecture – even ported to digital hardware – is probably not amenable to rapid recursive self-improvement.

Indeed, digitally emulating a human brain is not logically equivalent to "solving the human-level AGI problem", let alone to solving the problem of creating a Singularity-enabling self-improving AGI. But even so, there is a quite plausible-sounding path from

the former to the latter, such that from our present perspective, it seems very reasonable to surmise that having uploaded human brains to study would make the latter a heck of a lot easier.

An uploaded, digitized human brain is not really equivalent to a biological human brain, because it can be manipulated and studied a lot more flexibly and thoroughly. Of course, experimenting with sentient minds without their permission raises ethical questions; however, it seems likely that there will be no lack of willing volunteers, among software minds, for cognitive experimenation. Personally, as a mind scientist, if I were uploaded into a computer, I would happily volunteer copies of myself as subjects for non-painful experimentation. Currently, the legal systems of most countries do not allow individuals to volunteer themselves for scientific experiments; but it seems likely that these laws will change in a scenario where individuals can be replicated and reconstituted.

The ability to observe everything that happens inside an uploaded brain, and to make flexible manipulations and observe their consequences, is a huge difference from the situation with biological brains that have a disturbing tendency to break when you poke and prod them too much. For this reason, it seems highly likely to me that uploaded brains will – within years or at most decades, not centuries – lead us to a real science of human-level intelligence. If human-level AGI doesn't already exist by that point via other means, the knowledge gained from probing uploaded human brains will allow us to create it (thus, among many other more interesting things, providing the practical demonstration of computationalism that McDermott mentions).

Furthermore, it's worth noting that, even given merely human-level AGIs, Moore's law and related extrapolations suggest that the cost of human-level intelligence will then decrease by orders of magnitude per decade, a (very) significant economic impact. Relatedly, using human-level AGI scientists, an "artificial scientific community" could increase in scope exponentially at extremely mangeable cost. It seems quite plain that with this sort of rapidly expanded scientific capability, the step from digitally emulated human scientists to a principled understanding of human-level AGI should not be nearly so slow and difficult as the process of creating the first human-level AGI or human-brain emulation in the first place.

McDermott complains about Kurzweil's projection that

" 'Machines will be able to reformulate their own designs and augment their own capacities without limit'. Are there really principles of intelligence that would support limitless insight the way Bernoulli's principle supports the design of various kinds of lifting surfaces? It just seems to me that intelligence can't be boxed in that way."

and I agree that the "without limit" part seems an unnecessary speculation. Who knows what limitations may be discovered by vastly superhuman intelligences? The laws of physics and the properties of the physical universe impose their own limitations. But, I also note that a continual process of self-modification leading to unbounded increases in intelligence (within the constraints posed by the physical world) does not require universal "principles of intelligence" of the type McDermott mentions, and doubts. One can envision a series of ever-improving minds M_1, M_2, M_3, ..., where M_n understands enough to create a smarter M_{n+1} and not necessarily much more. No universal theory is

needed here. Different chains of ever-improving minds might discover different sorts of theories explaining different regions of mind-space. These are wild speculations, yes, but the point is that, contra McDermott, Kurzweilian "unlimited intelligence increase via progressive self-improvement" does not require any "boxing-in" of intelligence via universal laws.

A good analogy to invoke here, as pointed out by J. Storrs Hall (2006), is the scientific community itself – which is a continually self-improving collective intelligence composed internally of human-level intelligences. At each stage in its development, the scientific community knows enough to give rise to its next generation. Improvement may continue progressively, incrementally and perhaps unboundedly (within the constraints of the physical universe), without the need for any universal principles of scientific endeavor to be charted out in advance.

So, in spite of what McDermott says, there is no circular reasoning underlying the notion of Singularity approached via progressive self-improvement of human-level AGI systems. And Kurzweil does not commit the fallacy of assuming that emulating a human brain in a computer is logically equivalent to solving the AGI problem. Rather, Kurzweil correctly views human brain emulation as one route toward Singularity-enabling AGI, a route that involves:

1. Scanning human brains
2. Creating human brain emulations in advanced computer hardware
3. Studying these emulations to understand the principles underlying them
4. Creating various AGI systems embodying these principles, including AGI systems capable of radical self-modification and self-improvement

Kurzweil does not pose things in exactly this way but I believe it is a fair reading of *TSIN*. Appropriately (given the level of uncertainty associated with the relevant technologies), he does not make any commitments regarding how similar these Singularity-enabling, self-modifying AGI systems are going to be to the original human brain emulations.

4 Virtual-World Embodiment as an Alternate Plausible Route to Human-Level AGI

Screen capture of humanoid avatar w/ non-talking parrot in the Second Life virtual world

But what other plausible routes to AGI are there, aside from the brain-emulation route on which Kurzweil focuses most of his attention? Of course there are plenty, and since this is a brief essay I wish to avoid debating various contemporary paradigms on AGI design and their merits and shortcomings. My own detailed approach to AGI design, engineering and teaching can be found in references such as (Goertzel, 2006, 2007; Goertzel et al, 2004) if the reader is curious. What I want to discuss here is a general approach to developing and teaching AGI's and launching them into society that is quite different from what Kurzweil conceives in the context of whole-brain emulation.

The approach I'll discuss is based on virtual embodiment, and is interesting in the present context both because of what it suggests about the possibility of a non-Kurzweilian timeline for the development of human-level AGI, and because of the unique flavor of the Singularity scenarios to which it naturally leads.

The issue of the necessity for embodiment in AI is an old one, with great AI minds falling on both sides of the debate (the classic GOFAI systems are embodied only in a very limited sense; whereas Brooks (1999) and others have argued for real-world robotic embodiment as the golden path to AGI). My own view is somewhere in the middle: I think embodiment is very useful though probably not strictly necessary for AGI, and I think that at the present time, it may be more generally worthwhile for AI researchers to spend their time working with virtual embodiments in digital simulation worlds, rather than physical robots. Toward that end some of my current research involves connecting AI learning systems to virtual agents in the Second Life virtual world.

The notion of virtually embodied AI is nowhere near a new one, and can be traced back at least to Winograd's (1972) classic SHRDLU system. However, technology has advanced a long way since SHRDLU's day, and the power of virtual embodiment to assist AI is far greater in these days of Second Life, Word of Warcraft, HiPiHi, Creatures, Club Penguin and the like. To concretely understand the potential power of virtual embodiment for AGI, consider one potential project I've been considering undertaking during the next few years: a virtual talking parrot. Imagine millions of talking parrots spread across different online virtual worlds — all communicating in simple English. Each parrot has its own local memories, its own individual knowledge and habits and likes and dislikes — but there's also a common knowledge-base underlying all the parrots, which includes a common knowledge of English.

Next, suppose that an adaptive language learning algorithm is set up (based on one of the many available paradigms for such), so that the parrot-collective may continually improve its language understanding based on interactions with users. If things go well, then the parrots will get smarter and smarter at using language, as time goes on. And, of course, with better language capability, will come greater user appeal.

The idea of having an AI's brain filled up with linguistic knowledge via continual interaction with a vast number of humans, is very much in the spirit of the modern Web. Wikipedia is an obvious example of how the "wisdom of crowds" — when properly channeled — can result in impressive collective intelligence. Google is ultimately an even better example — the PageRank algorithm at the core of Google's technical success in search, is based on combining information from the Web links created by multi-millions of Website creators. And the intelligent targeted advertising engine that makes Google its billions of dollars is based on mining data created by the pointing and clicking behavior of the one billion Web users on the planet today. Like Wikipedia and Google, the mind of a talking-parrot tribe instructed by masses of virtual-world residents will embody knowledge implicit in the combination of many, many peoples' interactions with the parrots.

Another thing that's fascinating about virtual-world embodiment for language learning is the powerful possibilities it provides for disambiguation of linguistic constructs, and contextual learning of language rules. Michael Tomasello (2003), in his excellent book *Constructing a Language*, has given a very clear summary of the value of

social interaction and embodiment for language learning in human children. For a virtual parrot, the test of whether it has used English correctly, in a given instance, will come down to whether its human friends have rewarded it, and whether it has gotten what it wanted. If a parrot asks for food incoherently, it's less likely to get food — and since the virtual parrots will be programmed to want food, they will have motivation to learn to speak correctly. If a parrot interpret a human-controlled avatar's request "Fetch my hat please" incorrectly, then it won't get positive feedback from the avatar — and it will be programmed to want positive feedback.

The intersection between linguistic experience and embodied perceptual/active experience is one thing that makes the notion of a virtual talking parrot very fundamentally different from the "chatbots" on the Internet today. The other major difference, of course, is the presence of learning – chatbots as they currently exist rely almost entirely on hard-coded lists of expert rules. But the interest of many humans in interacting with chatbots suggests that virtual talking parrots or similar devices would be likely to meet with a large and enthusiastic audience.

Yes, humans interacting with parrots in virtual worlds can be expected to try to teach the parrots ridiculous things, obscene things, and so forth. But still, when it comes down to it, even pranksters and jokesters will have more fun with a parrot that can communicate better, and will prefer a parrot whose statements are comprehensible.

And of course parrots are not the end of the story. Once the collective wisdom of throngs of human teachers has induced powerful language understanding in the collective bird-brain, this language understanding (and the commonsense understanding coming along with it) will be useful for other purposes as well. Humanoid avatars — both human-baby avatars that may serve as more rewarding virtual companions than parrots or other virtual animals; and language-savvy human-adult avatars serving various useful and entertaining functions in online virtual worlds and games. Once AI's have learned enough that they can flexibly and adaptively explore online virtual worlds (and the Internet generally) and gather information according to their own goals using their linguistic facilities, it's easy to envision dramatic acceleration in their growth and understanding.

A baby AI has a lot of disadvantages compared to a baby human being: it lacks the intricate set of inductive biases built into the human brain, and it also lacks a set of teachers with a similar form and psyche to it … and for that matter, it lacks a really rich body and world. However, the presence of thousands to millions of teachers constitutes a large advantage for the AI over human babies. And a flexible AGI framework will be able to effectively exploit this advantage.

Now, how does this sort of vision relate to Kurzweil, the Singularity and all that? It's simply a very different pathway than human-brain emulation. Google doesn't emulate the human brain, yet in a sense it displays considerable intelligence. To demonstrate this, I just typed the query "How tall is a giraffe?" into Google. The snippet shown by the first search result reads:

Giraffe are the tallest animals in the world. Males can reach as high as 6 meters (19 feet) tall and female giraffe can measure up to5 meters -17 feet) tall. Their neck accounts for the added height and is by itself approximately 2.4 ...

This sort of experiment indicates that Google is a pretty decent natural-language question-answering system. Granted, it is not an incredibly smart question-answerer; for instance when I ask it "how tall is a giraffe that has been flattened by a steamroller?" its responses are irrelevant. But its responses to simple questions are often quite useful, due to the combination of the masses of information posted online, the inter-page links created by various Web authors, and the clever narrow-AI algorithms created by Google staff that make use of this text and these links. What we have here is considerable and useful artificial intelligence, achieved via the wisdom of crowds, combined with the ingenuity of the authors of Google's AI algorithms. Wikipedia is an interesting and related example: synergetically with Google or other search tools, it enhances the intelligence of the Web, in a way that specifically and cleverly leverages the collectivity of human intelligence.

It seems possible to harness the "wisdom of crowds" phenomenon underlying the Internet phenomena such as Google and Wikipedia for AGI, enabling AGI systems to learn from vast numbers of appropriately interacting human teachers. There are no proofs or guarantees about this sort of thing, but it does seem at least plausible that this sort of mechanism could lead to a dramatic acceleration in the intelligence of virtually-embodied AGI systems, and maybe even on a time-scale faster than brain-scanning and hardware advances lead to human-brain emulation. The human-brain-emulation route is better for drawing graphs – it forms a better "existence proof" – but personally I find the virtual-worlds approach more compelling as an AGI researcher, in part because it gives me something exciting to work on right now, instead of just sitting back and waiting for the brain-scanner and computer-hardware engineers.

Singularity-wise, one key difference between human-brain emulation and virtual-parrots-and-the-like has to do with the assumed level of integration between AI's and human society. Kurzweil already thinks that, by the time of the Singularity, humans will essentially be inseparable from the AI-incorporating technological substrate they have created. The virtual-agents pathway makes very concrete one way in which this integration might happen – in fact it might be the route by which AGI evolves in the first place. Right now many people consider themselves inseparable from their cellphones, search engines, and so forth. Suppose that in the Internet of 2015, Websites and word processors are largely replaced by some sort of 3D immersive reality – a superior Second Life with hardware and software support far beyond what exists now in 2007 – and that artificially intelligent agents are a critical part of this "metaversal"[1] Internet. Suppose that the AGI's involved in this metaverse become progressively more and more intelligent, year by year, due to their integration in the social network of human being interacting with them. When the AGI's reach human-level intelligence, they will be part of the human social network already. It won't be a matter of "us versus them"; in a sense it may be difficult to draw the line. Singularity-scenario-wise, this sort of path to AGI lends itself naturally to what I above called the "AI Big Brother" and "Singularity Steward" scenarios, in which AGI systems interact closely with human society to guide us through the future.

---

[1] See e.g. the Metaverse Roadmap, http://metaverseroadmap.org/

I don't care to pose an argument regarding the probability of the "virtual parrots" route versus the brain-emulation route or some other route. I consider both these routes, and others, highly plausible and worthy of serious consideration. The key point is that we can now explore these sorts of possibilities with dramatically more concreteness than we could do twenty or thirty years ago. As Kurzweil has adeptly argue, technology has advanced dramatically in ways that are critically relevant to AGI, and seems very likely to continue to do so -- and this is one of the key reasons I believe the time for human-level AGI really is near.

5 Conclusion

McDermott, in his critique of Kurzweil, levies the following complaint:

I wish he would stop writing these books! ... The field of AI is periodically plagued by 'hype hurricanes'.... Most of these storms are started or spun up by journalists, venture capitalists, or defense-department bureaucrats. But there are a very small number of people within the field who contribute to the problem, and Kurzweil is the worst offender.

My attitude could hardly be more opposite! Though I disagree with Ray Kurzweil on a number of points of emphasis, and a few points of substance, I believe he is doing pretty much exactly the right thing, by getting the general public enthused about the possibilities AGI is very likely going to bring us during the 21st century. I think the time is more than ripe for the mainstream of AI research to shift away from narrowly-constrained problem-solving and back to explicit pursuit of the general-intelligence goals that gave the field its name in the first place. In the last few years I have seen some signs that this may in fact be happening, and I find this quite encouraging.

The difficulty of predicting the future is not just a cliche', it's a basic fact of our existence. And part of the hypothesis of the Singularity is that this difficulty is just going to get worse and worse. One of my main critiques of Kurzweil is that he is simply too pat and confident in his predictions about the intrinsically unpredictable – quite differently from Vinge, who tends to emphasize the sheer unknowability and inscrutability of what's to come.

In spite of all this uncertainty, however, we must still plan our own actions, our own lives and our own research careers. My contention is that the hypothesis of the Singularity is a serious one, worthy of profound practical consideration; and even more emphatically, that the pursuit of human-level AGI deserves to be taken very seriously, at very least as much so as grand pursuits in other areas of science and engineering (gene therapy; building artificial organisms a la Craig Venter's work on *mycoplasma genitalium*; quantum computing; unifying physics; etc.). Yes, creating AGI is a big and difficult goal, but according to known science it is almost surely an achievable one. There are sound though not absolutely confident arguments that it may well be achievable within our lifetimes, via one of many plausible routes – and that its achievement may lead to remarkable things, for instance, one of the many Singularity scenarios discussed above, or other related possibilities not yet articulated.

References

- Anderson, J. R. (2000). Cognitive Psychology and Its Implications: Fifth Edition. New York: Worth Publishing.
- BBC News (2007). Mouse Brain Simulated on Computer, 27 April 2007. http://news.bbc.co.uk/2/hi/technology/6600965.stm
- Broderick, Damien (2003). Transcension. Tor Books.
- Brooks, Rodney (1999). Cambrian Intelligence. MIT Press.
- Cassimatis, N.L, E.K. Mueller, P.H. Winston (2006). Editors, Special Issue of AI Magazine on Achieving Human-Level Intelligence through Integrated Systems and Research. AI Magazine. Volume 27 Number 2.
- Crevier, Daniel (1993), AI: The Tumultuous Search for Artificial Intelligence, New York, NY: BasicBooks
- Drexler, K. Eric. (1992). Nanosystems: Molecular Machinery, Manufacturing, and Computation. New York: Wiley.
- Dreyfus, Hubert (1992). What Computers Still Can't Do. MIT Press.
- Franklin, Stan (2007). A foundational architecture for artificial general intelligence. In Advances in artificial general intelligence, Ed. by Ben Goertzel and Pei Wang:36-54. Amsterdam: IOS Press.
- Gilovich, Griffin and Kahneman (2002). Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press.
- Goertzel, Ben (2007). Virtual Easter Egg Hunting: A Thought-Experiment in Embodied Social Learning, Cognitive Process Integration, and the Dynamic Emergence of the Self. In Advances in artificial general intelligence, Ed. by Ben Goertzel and Pei Wang:36-54. Amsterdam: IOS Press.
- Goertzel, Ben (2006). Patterns, Hypergraphs and General Intelligence. Proceedings of International Joint Conference on Neural Networks, IJCNN 2006, Vancouver CA
- Goertzel, Ben, Moshe Looks and Cassio Pennachin (2004). Novamente: An Integrative Architecture for Artificial General Intelligence. Proceedings of AAAI Symposium on Achieving Human-Level Intelligence through Integrated Systems and Research, Washington DC, August 2004
- Goertzel, Ben and Stephan Bugaj (2006). The Path to Posthumanity. Academica Press.
- Goertzel, Ben and Cassio Pennachin, Eds. (2006). Artificial General Intelligence. Springer.
- Goertzel, Ben and Pei Wang, Eds. (2007). Advances in Artificial General Intelligence. IOS Press.
- Hall, J. Storrs (2007). Beyond AI: Creating the Conscience of the Machine. Prometheus Books.
- Hall, J. Storrs (2006). "Self-Improving AI: An Analysis"; lecture given at AI@50, Dartmouth, July 2006; journal version to appear in *Minds and Machines*
- Jones, R. M., & Wray, R. E. (2004). Toward an abstract machine architecture for intelligence. /Proceedings of the 2004 AAAI Workshop on Intelligent Agent

Architectures: Combining the Strengths of Software Engineering and Cognitive Systems/. For a complete list of publications regarding SOAR, see http://winter.eecs.umich.edu/soarwiki/Soar_Publications
- Kahane, Adam (2007). Solving Tough Problems. Berrett-Koehler
- Kurzweil, Ray (2005). The Singularity Is Near. Viking.
- McDermott, Drew (2006). Kurzweil's argument for the success of AI. Artif. Intell. 170(18): 1227-1233
- Minsky, Marvin (2006). The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. New York: Simon & Schuster.
- Thagard, P. (2005). Mind : Introduction to Cognitive Science. Cambridge, MA. MIT Press.
- Tomasello, Michael (2003). Constructing a A Language. Harvard University Press.
- Vinge, Vernor (1993). "The Coming Technological Singularity". Online at http://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html . The original version of this article was presented at the VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute, March 30-31, 1993. A slightly changed version appeared in the Winter 1993 issue of *Whole Earth Review*.
- Winograd, Terry (1972) . Understanding Natural Language. San Diego: Academic.
- Yudkowsky, Eliezer (2002). Creating Friendly AI. http://www.singinst.org/upload/CFAI.html
- Yudkowsky, Eliezer (2004). Coherent Extrapolated Volition. http://www.singinst.org/upload/CEV.html